DEPARTMENT OF THE AIR FORCE

AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT/GOR/MA/88D-3

A COMPARISON OF VARIABLE SELECTION CRITERIA FOR MULTIPLE LINEAR REGRESSION: A SIMULATION STUDY

THESIS

Ross J. Hansen
First Lieutenant, USAF

AFIT/GOR/MA/88D-3

Approved for public release; distribution unlimited

AFIT/GOR/MA/88D-3

A COMPARISON OF VARIABLE SELECTION CRITERIA FOR MULTIPLE LINEAR REGRESSION: A SIMULATION STUDY

THESIS

Presented to the Faculty of the School of Engineering

of the Air Force Institute of Technology

Air University

In Partial Fulfillment of the

Requirements for the Degree of

Master of Science in Operations Research

Ross J. Hansen, B.S.

First Lieutenant USAF

December 1988

Acknowledgements

Before starting my thesis, my thesis advisor, Dr. David R. Barr, told me "the most difficult paper a professor ever has to write is his advisee's." Well, hopefully this one was not too difficult.

I would like to express my sincerest gratitude to Dr. Barr for his expertise, support, and especially for his uncanny ability to see the "forest for the trees."

I would like to thank Major W. Kenneth Bauer who assisted me in Response Surface Methodology and regression theory, and Dr. Joseph P. Cain who acted as my reader on my thesis committee.

I would also like to thank my wife, [redacted] for her understanding and encouragement.

Ross J. Hansen

Table of Contents

List of Figures

List of Tables

Abstract

The purpose of this thesis was to identify three promising least squares selection procedures discussed in the literature during the previous decade and then test them using simulation. The three criteria chosen for this study were minimum mean square error (Min MSE), minimum S_p, and minimum C_p.

Most of the previous simulations in this area are limited to investigating the usefulness of variable selection criteria when all relevant regressors and some noise variables are available. It is questionable whether all relevant variables will be included. This research has examined the effects of not including a significant variable in the variable pool.

In examining each criterion, emphasis was placed on the technique's performance under varying amounts of multicollinearity, variable variation, number of variables, and sample size. Response Surface Methodology was used to determine the effects of varying these factors. A comparison was then made using the results from the Response Surface.

To supplement the simulation research a comprehensive literature review of the most current journal articles

dealing with several least squares criteria has been provided. This review includes a discussion of each technique's strengths and weaknesses. Since many of the least squares variable selection criteria are addressed, this thesis serves as a useful starting place for various regression questions.

A COMPARISON OF VARIABLE SELECTION CRITERIA FOR MULTIPLE LINEAR REGRESSION: A SIMULATION STUDY

I. Introduction

Background

Linear regression is a statistical tool used to fit data to a surface. From the surface, predicted values for the dependent variable are determined with confidence. Unfortunately, a difficult aspect of regression analysis is determining the best set of independent variables to include in the linear regression model.

Before the wide use of computers, the analyst was forced to rely on his intuition for determining which variables should be included in the model. Variable selection techniques were available but difficult to implement. It was important for the analyst to "screen" the variables first by a reasonability test. For example, say the dependent variable is the height of a man. Possible independent variables are: size of hand, size of feet, and blood-type. It seems reasonable to assume that the height of a man is dependent on the size of his hands and feet. But, it does not make sense to assume his blood-type significantly contributes to his height. Therefore, the

analyst would include only feet and hand size into his variable pool. The analyst could perform his analysis on all the variables in the variable pool, which is ordinary least squares (OLS), or he could use a variable selection method to find the most significant variables. If he simply calculates the regression equation containing both feet size and hand size, the method is called OLS. A better model might include only one of these variables, so the analyst might want to use one of the selection criteria on the variable pool.

Due to computer limitations, only a few criteria for selecting variables were available, and the literature out-paced implementation. With computer advancements, implementation of new techniques should be increasing; however, this is not necessarily true. Often, these published techniques are forgotten. Rarely are the "new techniques" tested and implemented. In fact, most of the techniques currently in use are 1970 vintage, or earlier. To be sure, some of these "new techniques" are superior to those in use. Since these criteria are lost in the literature, research may be suffering.

Several problems can occur if the variable selection is left solely to the computer. First, there are literally hundreds of variable selection criteria available. Under certain circumstances, one method may be superior to another. By mindlessly using a technique, the resulting

model might be worthless. Certainly there are methods which "screen" better than others. However, there is not a method which absolutely picks only true variables. During a simulation study, Flack and Chang demonstrated that when the variable pool contains true variables, those which significantly contribute to the dependent variable, and random noise variables, random noise variables were often chosen (8:85). Only under the most ideal circumstance were a fair amount of true variables chosen.

The problem of selecting noise terms was further illustrated by Freedman. (9) By constructing simulated data containing only noise variables, Freedman demonstrated that high R^2 values could result when a model contained variables which are theoretically independent of the dependent variable. (9:153) The point being, selection criteria are not fail proof. The model chosen could possibly contain several noise variables. The more noise variables in the variable pool (those not screened by other means), the more likely one or more will be included in the model.

A subsequent simulation study performed by Hoerl, Hoerl, and Schuenemeyer identified and tested the effects of sample size, random noise variables and correlated data on biased regression techniques and least squares regression. (13) Their results indicated that both biased and least squares regression performed equally well except for stepwise

regression and principle component regression which performed poorly. (13:369)

Objective

The objective of this research is to identify three promising least squares selection procedures discussed in the literature during the previous decade and then test them using simulation. The simulation was similar to Flack and Chang's in the sense that the effects of noise variables and sample size were examined (8:84-86), but did not make the same assumptions and used different criteria. Flack and Chang tested the R^2 and stepwise procedures. Like several other simulations, they included all the true variables and some noise variables and then determined how many noise variables were selected. This research examined three criteria, minimum mean square error (Min MSE), minimum S_p, and minimum C_p and included random noise variables. An extension of Flack and Chang's work has been made. Flack and Chang's research did not examine the problem of not including all the significant variables. In the previous example, data on the height, size of feet, size of hands, and blood-type of a man was collected. It is possible that the weight of a man is a also a significant factor in determining the man's height, but that no information has been collected on the man's weight. Most criteria are based on the assumption that all relevant variables are included in the variable pool; however, the fact is, all relevant

variables may not be included. This research has examined the effects of not including a significant variable in the variable pool.

In examining each criterion, emphasis was placed on the technique's performance under varying amounts of multicollinerarity, variable variation, number of variables, and sample size. Response Surface Methodology was used to determine the effects of varying these factors. A comparison was then made using the results from the Response Surface.

Previous simulations have based their comparisons solely on the number of noise variables chosen. (2;8;12) The weakness of this performance measure lies in the fact that as the number of variables in the variable pool changes, the likelihood of choosing solely the correct variables also changes. In this study an alternative performance measure was be used for comparison.

To supplement the simulation research a comprehensive literature review of the most current journal articles dealing with several least squares criteria is provided. This review includes a discussion of each technique's strengths and weaknesses. Since many of the least squares variable selection criteria are addressed, this thesis serves as a useful starting point for various regression questions.

II. Concept Overview

Least Squares Regression

Assumptions. Assumptions are made prior to constructing a least squares linear regression. First, assume the collected data represents the population it came from. That is, the data reflects the normal case of the variable. Second, assume the error terms are independent and identically distributed, from a normal distribution, with a mean of zero and variance σ^2.

Notation. The goal in linear regression is to find the "best-subset" of independent variables to include in the model which adequately predicts the value of the dependent variable. In general, the linear least squares regression equation is written in the following manner:

$$Y = B_0 + B_1X_1 + B_2X_2 + \ldots + B_kX_k + \epsilon \qquad (1)$$

where Y is the observed value of the independent variable.

B_0 is the constant term.

$B_1, B_2, \ldots, B_k$ are the constant terms for the dependent variables $X_1, X_2, \ldots, X_k$.

k is the number of independent variables included in the model.

ϵ is the error term.

If there are n observations, or data points, the above equation may be written as:

$$\Sigma\, Y_i = \Sigma\, (B_{0i} + B_{1i}X_{1i} + B_{2i}X_{2i} + \ldots + B_{ki}X_{ki} + \epsilon_i) \qquad (2)$$

For convenience, the above equation can be written in matrix notation.

$$\mathbf{Y} = \mathbf{X}\,\mathbf{B} + \boldsymbol{\epsilon} \qquad (3)$$

where **Y** is a (n x 1) column vector:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix}$$

and **X** is a (n x (k+1)) matrix.

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & . & . & . & X_{1k} \\ 1 & X_{21} & X_{22} & . & . & . & X_{2k} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 1 & X_{n1} & X_{n2} & . & . & . & X_{nk} \end{bmatrix}$$

The first column contains all ones for the constant terms. The remaining columns contain the X_{ij} independent variables. The **X** matrix is commonly referred to as the design matrix. **B** is a (k x 1) column vector:

$$\mathbf{B} = \begin{bmatrix} B_0 \\ B_1 \\ \cdot \\ \cdot \\ \cdot \\ B_k \end{bmatrix}$$

and ϵ is a (n x 1) column vector:

$$\epsilon = \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{bmatrix}$$

In least squares regression, each subset of regression variables generates a surface which minimizes the squared distance (error) between the observed values for the dependent variables, Y, and the predicted values for the dependent variable $\hat{Y}$.

$$\min \Sigma \epsilon^2 = \min \Sigma (\hat{Y}_i - Y_i)^2 \qquad (4)$$

The goal is to find the subset of variables which minimizes the squared distances between the actual values observed and the fitted surface. The sum of the squared-error values is commonly referred to as the sum-of-squares error (SSE).

Graphically, in two dimensions a regression resembles the following:

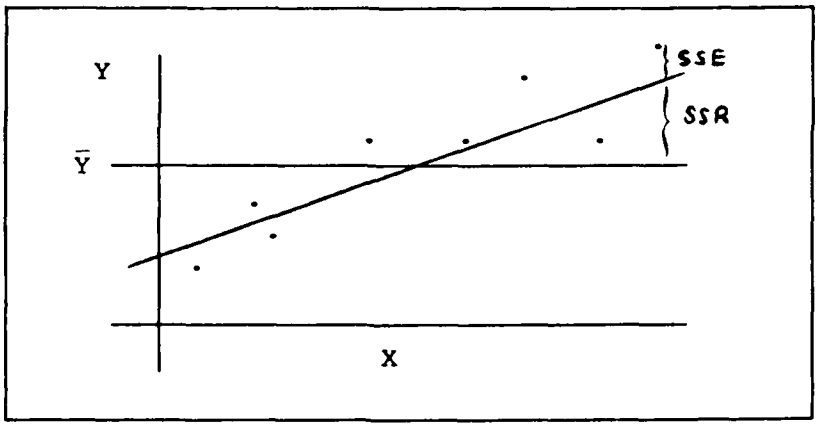

Figure 1. Two-dimensional representation of linear least squares regression

where SSR is the sum of the squared distances from the mean to the regression line and is called Regression Sum of Squares.

SSE is the sum of the squared distances from the point to the regression line and is called Sum of Squares Error.

SST is the sum of SSR and SSE and called Sum of Squares Total.

$$SST = SSR + SSE \quad (5)$$

III. Review of Past Research

Scope

The main focus of this literature search is limited to variable selection techniques published since 1978. This limitation was chosen for three reasons. First, two GOR students, Captain Larry J. Pulcher (24:1-137), and Second Lieutenant Joseph R. Cafarella, Jr. (4:1-127), performed literature reviews on variable selection techniques for their theses, written in 1978 and 1979 respectively. Second, most of the implemented techniques were published prior to 1978. Third, an excellent review of the pre-1980 techniques can be found in Hocking's 1976 paper (12).

Method of Treatment and Organization

This literature review reports methods for selecting the number of variables to include in a linear regression equation, as well as specific techniques for selecting the "best-subset" of variables to include in the regression equation.

Review

If the true variance, σ^2, of the independent variables is known, selecting the proper regression model is relatively simple (7:294). However, the true variance is rarely known *a priori*. It is more likely that the underlying variance will have to be estimated. Therefore, techniques must be

available for the analyst so that he may properly select the variables to include in a linear regression equation.

Number of Variables. Draper and Smith (7:294) discuss two opposing viewpoints concerning the proper number of independent variables (or regressors) to include in a linear regression equation. One viewpoint is all variables should be included for both a reliable fit and predictive purposes. The opposite point-of-view is to include as few variables as possible, which adequately predict, in the model. The latter point-of-view is commonly referred to as parsimony.

For predictive purposes, Bayesian statisticians recommend including all possible variables (28:1553). They argue the cost associated with collecting and maintaining the data must be considered. If these costs are zero, than every variable should be included. Only by including as many variables as possible can one be certain to obtain the best prediction.

Cafarella (4:14) points out the major flaw of including every possible variable. The resulting model over-fits the data. That is, the independent variables not only predict the dependent variable, but also predict the variation, or noise. The resulting equation is useful for predicting the past, but not very useful for predicting the future. Thus, it is arguable whether including all possible regressors is the best course of action.

Even the Bayesian statisticians agree that if the costs of collecting and maintaining data are high, then one should include as few variables as possible in the model. However, some believe in the concept of parsimony regardless of the costs. Parsimony is predicting as much as possible, with as little as possible. Trader states, "Variable selection techniques, which facilitate parsimony, have been viewed primarily as providing approximate representations of a true underlying process." (28:1553) The advantages of parsimony is that model does not over-fit the data, and it does not cost as much as a larger model to collect and maintain. However, the question remains: does the model contain enough variables to predict adequately?

Recent literature (23:509-16; 3:131-6; 24:45-54) suggests a method for determining the optimal number of regressors for a linear regression equation. The two driving factors for determining the optimal number of regressors are sample size and the mean square error of prediction (MSEP). (23:514) MSEP is the squared difference between the predicted value, $\hat{Y}_{n+1}$, and the actual value, Y_{n+1}, divided by the degrees of freedom. The MSEP has been decomposed by Breiman and Freedman (3:131) into the conditional MSEP, and the unconditional MSEP. The conditional MSEP is

$$M = M_{np} = \Sigma\{(\hat{Y}_{n+1} - Y_{n+1})^2 \;/\; Y_i \text{ and } X_{ij} \text{ for all } j \text{ and } i = 1,\ldots,n\} \qquad (6)$$

while the unconditional MSEP is defined as

$$U = U_{np} = E[M_{np}] \tag{7}$$

With the mean square error of prediction and a given sample size, one can find the optimal number of parameters, p^*, to include in the model.

$$p^* = k + 1 \tag{8}$$

where

p^* is the number of parameters in the model with Min MSEP.

k is the number of independent variables.

When either M or U from Eq (5) and Eq (6) are minimized, the corresponding p^* is the optimal number of parameters for the regression model. It is apparent that when the sample size is small, the optimal number of parameters tends to be small (28:1562).

Choosing the Best Subset of Variables. Regardless of whether the optimal number of parameters has been determined, a criterion must be used to select which variables should be included in the regression equation. Miller separates possible criteria into "(i) those which guarantee to find the best fitting subset of some of all sizes, and (ii) the `cheap' methods which sometimes find the best-fitting subsets." (20:391)

"Guaranteed Methods". Miller specifically mentions two methods which guarantee choosing the best-fitting

variables. These two methods are all-subset and branch-and-bound.

All-Subsets. In the 1984 article, "Selection of Subsets of Regression Variables" (20:391), Miller claims that only by an exhaustive search of all 2^k-1 possible subsets can one be assured of finding the best-fitting model. He further suggests the all-subset procedure is practical only when the number of possible variables is less than twenty. With the improvement of computers over the last five years, it might be feasible to test every subset for more than twenty variables. However, the necessary number of regressions increases rapidly. For instance, with twenty variables, over one million subsets must be fit. By increasing the number of variables by just five to twenty-five, the number of necessary regressions increases to roughly 33.5 million.

The all-subset procedure is not without criticism. One of the underlying assumptions in regression is the variable pool contains all the relevant variables and some extraneous variables. (21:160) However, it is possible some of the relevant variables are not included in the variable pool. Because of this, the all-subset procedure may not find the best model in population terms, but always finds the best model within the variable pool (2:3). In fact, backward elimination, (one of the "cheap methods," which will be discussed later), does slightly better than all-subsets

under certain circumstances. Although Berk criticizes the all-subset procedure under certain circumstances, he notes that it usually does well (2:3). With large sample sizes (approaching the population size) all-subset is the best.

Branch-and-Bound. Miller includes the branch-and-bound technique as one of the guaranteed methods when the optimal number of variables has been determined (20:391). The branch-and-bound technique computes only a fraction of the possible models. It eliminates subsets from consideration if a calculated subset fits better than the eliminated subset could possibly obtain. The eliminated subsets are said to be "dominated." With some of the models "dominated," the all-possible subset procedure can be implemented on the subsets that are not eliminated by branch-and-bound. The branch-and-bound technique is not only useful in eliminating possible subsets, but can be modified to be an efficient all-subset procedure (10:510). The necessary modification requires that the "testing" mechanism be turned off.

"Cheap-Methods." Miller conveniently labels all techniques other than the all-subset and the branch-and-bound procedures as "cheap." These techniques are cheap in the sense that they do not require as many calculations for finding a solution. Consequently, none of the cheap-methods are guaranteed to find the best-fitting model.

Backward Elimination. Backward elimination is a technique which considers all of the possible variables in the original model, then eliminates, one at a time, those variables which do not significantly contribute to the regression. According to Draper and Smith, "this [backward elimination] is a satisfactory procedure, especially for statisticians who like to see all the variables in the equation once in order `not to miss anything'." (7:307) This procedure is satisfactory if the data is not highly correlated.

Nonetheless, backward elimination has its drawbacks. One major drawback is once a variable is eliminated, it is not considered again. It is possible that an eliminated variable, which did not look significant when it was eliminated, will look significant after other variables are eliminated. Another criticism of backward elimination is that it is not useable when the number of possible variables exceeds the number of data points. Due to the matrix manipulations necessary in regression, this procedure cannot be used when variables exceed data points.

Forward Selection. Forward Selection is similar to backward elimination. However, instead of all of the variables being included in the original model, then eliminating one variable at a time, forward selection starts with one variable, and includes a variable-at-a-time. The major advantage of this technique is that it can be

implemented when the number of possible variables exceeds the sample size. This technique can be used since the original model has one variable, to which variables are subsequently added. (20:392).

The disadvantages of forward selection are similar to backward elimination. Like backward elimination, forward selection performs poorly when variables are highly correlated. Also, once a variable enters the regression equation, it is no longer considered for elimination.

Stepwise Regression. Stepwise regression is a procedure which combines both the forward and backward selection procedures. Like forward selection, variables are selected to enter one at a time. However, after a variable enters the regression equation, all variables in the subset are reconsidered for elimination. This improvement allows for the deletion of those variables which do not significantly contribute to the model.

Draper and Smith recommend this technique. However, they add the caveat that it is not be used blindly (7:310). If stepwise regression is used with little thought, many random noise variables can be selected. As is the case with all of the techniques, the likelihood of selecting a noise variable is fairly high if the number of noise variables is high.

To combat this problem Miller suggests an alternative stopping criterion for stepwise regression. His suggestion is to add "known" extraneous variables to the variable pool

(randomly generated noise) then execute the stepwise regression procedure. Once one of the "known" extraneous variables is selected, terminate the selection process. (20:395) The concept behind his suggestion is that once a "known" extraneous variable (one that was augmented to the variable pool) is selected, there is no more useful information to be gained by adding another variable. The model selected will include all variables picked not including the known extraneous variable. For the technique to be useful, Miller recommends that the number of added extraneous variables should be approximately equal to number of variables already in the variable pool. (20:395)

Excluding its weakness when several extraneous variables are in the variable pool, stepwise regression performs well when there is little correlation. Many computer packages use a significance level of 0.05 for forward, backward, and stepwise procedure. The choice of 0.05 seems to be arbitrary and unjustified. In fact, the results of a simulation study conducted by Hoerl, Hoerl, and Schuenemeyer indicate that a significance level of 0.05 often leads to poor results (13:378). They demonstrated that many of the true variables are not selected when the significance level is 0.05. However, most variables are picked using a significance level between 0.15 and 0.25

Near-Optimal-Model. Narula and Wellington (21:169) propose the near-best-model for Mean Square

Absolute Errors (MSAE) in 1983. MSAE is the sum of the absolute difference between the predicted value of Y and the actual value of Y, divided by the degrees of freedom.

$$\frac{\Sigma \mid \hat{Y}_i - Y_i \mid}{(n - p)} \tag{9}$$

where

n is the sample size.

p is the number of parameters in the model.

$\hat{Y}$ is the predicted value for Y.

Y is the actual value of Y.

Narula and Wellington put a "lower and an upper limit on the number of variables to be included in an effort to accelerate the implicit enumeration algorithms and aid the investigator in selecting a model." (21:169) Although they apply the near-optimal technique to MSAE, they suggest it may be effective in the least squares case.

Similar to Narula and Wellington's proposed near-best-model for Mean Square Absolute Errors, Huang and Panchapakesan have implemented a procedure to eliminate inferior models with some guaranteed probability (14:753). Their procedure is based on the residual sums of squares and is similar to Furnival and Wilson's branch-and-bound technique. The near-best-model for Mean Square Absolute Errors is based on SSR making it easy to implement. What makes this technique especially appealing is that once a

model is rejected, then all combinations of its variables are also eliminated. (14:758).

An example of how the near-best-method works is the following:

Suppose the subset of variables under consideration is $\{X_1, X_2, X_3\}$. If this subset of variables is rejected, then all subsets of X_1, X_2, and X_3 would also be rejected (e.g. $\{X_1\}$, $\{X_2\}$, $\{X_3\}$, $\{X_1, X_2\}$, etc.).

Mallows C_p. Mallows C_p is a statistic used to determine the best model when the independent variables are fixed. C_p is an approximation of MSEP.

$$C_p = \frac{SSR}{s^2} + 2p - n \qquad (10)$$

where SSR is the Regression Sum of Squares

s^2 is the estimate for the variance

p is the number of parameters

n is the number of data points

Theoretically the value of C_p is p. Therefore, when C_p is approximately equal to p, the model is good. Draper and Smith suggest using this criterion in conjunction with stepwise regression to obtain the best subset (7:341). It should be noted, however, as the variance approaches zero, the C_p statistic can not be calculated. Therefore this method has limitations especially when the fit is perfect.

Barr pointed out a weakness of Mallows C_p. Since s^2, used in the C_p statistic, is estimated from the original variable pool, it could be biased and larger than the true variance. (1:5) If this is case, the C_p statistic will be deflated causing the wrong model to be selected.

A limitation of C_p, as well as many other statistics, is that it "depend[s] on the observed data only through sufficient statistics, so they model average behavior of the fit of a model to the data." (30:27) Weisberg developed a procedure which allocates the C_p statistic to individual cases. The advantage of Weisberg's procedure is if the model under consideration is biased, it provides a means to determine the bias of using a subset model instead of the entire model (30:28)

Another application of the C_p statistic is to choose the model which has the smallest C_p value. (15:863) By choosing the model with the minimum C_p, it is believed that one is choosing the model with the minimum prediction error. This is appealing, especially when it is difficult to determine the optimal subset using the C_p close to p criterion. Since the Min C_p criterion is based on minimum prediction error it is based on a sound principle. However, like the C_p close to p criterion, Min C_p is derived under the assumption that the independent variables are fixed. Since this rarely happens in practice, there is some question to the usefulness of the Min C_p criterion. Judge,

Griffiths, Carter, Lutkepohl, and Lee recommend that the Min C_p procedure should not be used in any applied work. (15:864)

Coefficient of Determination. The coefficient of determination, R^2, is a statistic which gives an estimation of the amount of variation about the mean which is explained by the model.

$$R^2 = \frac{\Sigma (\hat{Y}_i - \bar{Y})^2}{\Sigma (Y_i - \bar{Y})^2} \qquad (11)$$

where

$\hat{Y}_i$ is the predicted value of Y_i.

Y_i is the actual value of Y_i.

$\bar{Y}$ is the mean of Y.

At first one might believe that it is desirable to find the model which has the maximum R^2, since it explains the most variation about the mean. However, this is not necessarily the best true. Certainly when we look at the R^2 value we would like to see a large value, but it should not be used as the only measure for subset selection. Maximum R^2 receives little praise as far as its usefulness in determining a good fit. The major pitfall of using R^2 is that whenever a variable is added, it will increase R^2. R^2 will increase regardless of whether the variable has anything to do with the dependent variable. According to Healy 1986, "In particular, the multiple correlation coefficient is not really a regression-related concept at

all. It is basically defined to be the largest possible correlation between the y-variate and any linear function of the x's and this only makes sense when y and x's have a joint probability distribution." (11:1984) If maximum R^2 is used as the selection criterion, the model containing all variables will always be selected.

Maximum Adjusted R^2 or Minimum MSE. For simplicity only maximum Adjusted R^2 will be discussed. However, maximum Adjusted R^2 and Minimum MSE test exactly the same thing.

Adjusted R^2 is related to R^2, but an adjustment has been made for the degrees of freedom. The following equation shows the relationship between R^2 and Adjusted R^2.

$$\text{Adjusted } R^2 = 1 - \frac{(1-R^2)(n-1)}{(n-p)} \qquad (12)$$

According to Draper and Smith, the adjusted R^2 statistic can be used not only to compare models for the same data set (the same variable selection discussed in all other sections of this literature review), but also to compare models taken from two entirely different data sets (7:92). However, they do not recommend using the Adjusted R^2 statistic in the latter role.

The Adjusted R^2 statistic (or the minimum MSE criterion) is still widely used in practice.

S_p. The S_p criterion, originally proposed by Hocking in 1976 (12:20), has considerable appeal and

consequently receives praise in recent years. The S_p statistic is an approximation of the MSEP based solely on the data and number of variables. As is the case with MSEP, the goal of this criterion is to find the minimum value.

$$S_p = \frac{SSE}{(n-p)(n-p-2)} \tag{13}$$

Breiman and Freedman point out that the S_p statistic does not necessary provide an accurate approximation of MSEP, but works none the less. (3:132)

The advantages of this method are numerous. Looking at Eq (12) gives the reader an idea of the relative ease with which S_p is calculated. What makes S_p even more appealing is it is based on MSEP. Thompson points out, "This method [S_p] is based on a sound criterion - that of minimizing the expected squared distance between the true and predicted values of the dependent variable, y." (27:6) Since S_p is an approximation of MSEP, it can be used like MSEP to determine the optimal number of regressors to include in the model. (3:132)

S_p is not without its disadvantages. It must be calculated for all 2^k -1 possible subsets. (27:6) Even though it requires relatively little computation effort, it does require that many regressions be run. Through counter examples Brieman and Freedman show that when true variance due to prediction equals zero, the S_p criterion fails to

pick the optimal number of variables to include in the model. (3:132)

Conclusion

A considerable amount of literature has been devoted to linear regression in the last decade. This review has covered a few of the techniques mentioned in the literature. By no means should this review be considered exhaustive. However, it is intended to provide an excellent review of many least squares regression techniques from the last decade.

Since the calculations required for the "guaranteed" all-subsets method virtually make it impossible to use, much of the effort in the field is to find techniques which require far less calculation. Unfortunately, the savings in computer time is off-set by the results. None of the techniques mentioned in the "cheap" section of this report guarantee the best-fitting subset of regressors will be chosen.

Biased Regressors

Up to this point the discussion of literature has been restricted to selection techniques for least squares. The advantage of least squares over the techniques discussed in this section is that the estimators are best linear unbiased estimators, or BLUE. However, under certain circumstances

the estimators from least squares can be inferior to biased regressors.

An estimator is said to be unbiased if its expected value of the estimator is the parameter value itself.

$$E(b) = B \tag{14}$$

Conversely, if the expected value of the estimator is not equal to the parameter value, it is said to be biased.

The following techniques are intended to provide the reader with some insight into when the use of biased estimators may be a better course of action than least squares. Rather than discuss several biased techniques, this discussion will be limited to two methods which have received a great deal of attention in recent literature. Both ridge regression and principle components, the two methods to be discussed, are methods for dealing with multicollinearity.

Ridge Regression. There has been a significant amount of attention given to Ridge Regression in the past decade. Ridge Regression is a procedure which attempts to overcome the effects of highly correlated data. As mentioned before, many of the least squares selection criteria perform poorly under such conditions.

There are two situations for which Draper and Smith "wholeheartedly recommend ridge regression." The first situation is when there is "a Bayesian formulation of a regression problem with specific prior knowledge of a

certain type on the parameters." (7:319) The advantage over least squares in this situation is that the B's are restricted a priori. The second situation in which Draper and Smith recommend Ridge Regression is "a formulation of a regression problem as one of least squares subject to a specific type of restriction on the parameters." (7:320) Draper and Smith comment that under the two previous situations ridge regression is absolutely the correct model to use. However, they also point out that under any other situation, there needs to be further investigation to determine the usefulness of ridge regression.

Dempster, Schatzoff, and Wermuth performed a simulation study on alternatives to least squares regression in 1977. They believe ridge regression offers a drastic improvement over least squares regression when there is a great deal of correlation between the independent variables (5:77). An interesting comment by the authors is, "From a frequentist standpoint, it has long been recognized that good mean squared error properties do not necessarily follow from the celebrated minimum variance unbiasedness properties of least squares, since in certain regions of the parameter space the loss from increasing the squared bias can be overcompensated by reducing variance." (5:77)

In a subsequent study (19), Makin tested the ridge regression procedure and found it performed well. Even

though his models were fairly trivial, he concluded that ridge regression is, in fact, a good technique.

It should not, however, be misconstrued that ridge regression is a cure-all. If many of the variables under consideration are not related to the dependent variable ridge regression does not perform very well (13:375). A suggestion to counter this drawback is to delete a portion of the variables if there is reason to believe many have little to do with the independent variable. Another caveat of using ridge regression is to recognize there is an entire family of ridge regression techniques. One in particular, REGF, is not recommended by Dempster, Schatzoff, Wermuth as a practical tool.

Principle Components. *Principle components is another* technique used to combat collinearity and has received a considerable amount of attention recently . Collinearity tends to yield highly unstable estimates because the design matrix is nearly singular. The goal of principle components is similar to ridge regression in the sense that the goal is to produce estimates which are not greatly effected by collinear data. However, unlike ridge regression which assumes nonsample information (15:909), principle components does not assume any nonsample information is available.

In principle components, only subsets of the variables are considered. The choice of subsets can be made "by economic theory, previous statistical results, or ad hoc

dimensionality reduction procedures." (15:909) By reducing the number of variables under consideration, principle components attempts to eliminate variables which do not necessarily contribute to the model and cause unnecessary collinearity.

There is some criticism of the method of principle components. One criticism by Judge, Griffiths, Hill, Lutkepohl, and Lee, is that this method relies on the data in hand, which is the same weaknesses of every search method. It is not guaranteed to find the best subset of regressors to include in the model (15:910). It is quite possible that the subset of variables chosen does not contain all of the true regressors. In their simulation study, Hoerl, Hoerl, and Schuenemeyer demonstrated that principle components performs poorly. (13:375) "It was rarely noticeably better than LS [least squares], often worse than LS and uniformly inferior to RRB [a ridge regression technique] for all criteria." (13:375)

Conclusion

If a situation arises when significant multicollinearity exists, and it is not appropriate to use least squares, it may be appropriate to use a biased regression technique. Two of the methods which have received recent attention are ridge regression and principle components. However, like least squares techniques, these techniques give guarantees for finding the best model. Also like least squares, both

of these techniques require certain conditions to perform optimally. Ridge regression requires some type of nonsample information while principle components does not. However, principle components does not perform nearly as well as ridge regression and even least squares. These trade-offs must be considered when deciding on a technique to combat collinearity.

IV. Methodology and Model Development

The three techniques selected for this research were minimum MSE, minimum C_p, and minimum S_p.These techniques were chosen for the following reasons:

(1) Each of these techniques are absolute criterion. That is, they is no "art" necessary. The S_p, C_p and MSE statistic for each subset of variables is obtainable under some all-subset procedures, such as PROC R^2 on SAS. One only needs to find the minimum value for each technique and the corresponding set of variables to implement these techniques. Contrasted with the C_p close to p technique which leaves confusion as to whether a model is superior to another simply on the basis of the difference between C_p and p.

(2) All three techniques appear in the last decade's literature. The Minimum MSE procedure used to be one of the most widely used methods. Its appeal over techniques such as Max R^2 stems from its adjustment for degrees of freedom. More recently, S_p seems has become the most popular technique. Its appeal is based on the principle of minimizing mean square errors of prediction.

(3) The C_p criterion is also based on MSEP, and some of authors praise this criterion.

Overview

With any procedure, certain assumptions must hold to insure proper execution and least squares is no exception. As mentioned, by using least squares one assumes the data is typical of the population, and the error terms are independent and identically distributed from a normal population with an expected value of zero and a constant variance σ^2. Another assumption often made, yet not entirely justified, is that all the variables collected relate to the dependent variable.

Objective

The goal of this thesis is to determine the behavior of the Min MSE, Min C_p, and Min S_p variable selection criteria. To do this, response surface techniques was implemented.

Response Surface Methodology

When selecting variables to include in the original variable pool for the least squares regression equation, it is desirable to select only those variables which significantly contribute to the model. However, there is no guaranteed method to screen the extraneous variables(random noise terms which do not contribute at all to the model) from the variable pool. Even more discouraging, once in the variable pool there is no criterion which guarantees that no extraneous variables will be chosen for the model. The problem of selecting extraneous variables is not restricted to the "cheap" methods. In fact, one of Miller's guaranteed

methods, the all-subset procedure, which is suppose to perform quite well, also occasionally chooses extraneous variables. The limitations of variable selection criteria, including the "guaranteed" techniques will be emphasized since all three of the criteria under consideration are all-subset procedures.

Since there really are no "guaranteed methods" for capturing all and only the true variables, it is advantageous to strive for the highest percentage of those variables chosen correctly. Therefore the performance measure for the Response Surface is the following:

$$PM = \frac{\text{(number of correct variables chosen)}}{\text{(number chosen)}} \qquad (15)$$

PM is the logical choice for two reasons. First, the best model may not include <u>all</u> variables it is generated from, but only the most significant. Even though an independent variable may have been generated from ten variables, the best model of reality using, say MSE, may only contain two of those variables. Therefore, PM compensates by determining the percentage of correct variables chosen. Second, PM is influenced by the number of extraneous variables chosen. If only two variables are chosen, and one of them is an extraneous variable, it is worse than when ten

variables are chosen and only one of them is extraneous. PM allows for such an adjustment.

By using Response Surface Methodology, equations made up of significant factors (multicollinearity, sample size, etc.) for each of the three criteria can be fitted. Using these equations, a comparison of all three criteria under best-case and worst-case scenarios can be completed.

Data Generation For Response Surface. The data for this study was generated from the following equation:

$$Y_i = X_{1i} + X_{2i} + X_{3i} + X_{4i} + \epsilon_i \qquad (16)$$

where

Y_i is the dependent variable.

$X_{1i},\ldots,X_{4i}$ are correlated, randomly generated dependent variables.

ϵ_i is a noise term to create variance in the model.

Most simulation studies investigate variable selection techniques with all relevant dependent variables plus some extraneous variables included in the variable pool. This study attempts to show what happens when one of the relevant variables is not included in the variable pool.

After the data is created from equation (16), the independent variable, X_4, is dropped from consideration. This simulates the situation which arises when a significant variable is not included in the variable pool. In addition

either one or three noise variables are included in the variable pool to simulate data collection on extraneous variables.

Factors. An equation made up of significant factors and factor interactions which adequately predicts the usefulness of each of the three techniques must be found. Ideally, the significant factors are observable a priori (e.g. sample size, correlation, etc.), thus, allowing for compensation prior to use. The following are the factors for this study:

(1) The number of extraneous variables (EX_1, EX_2, EX_3) in the original variable pool. These variables are noise, and therefore are theoretically independent from the dependent variable. In this study, at the low setting the number of extraneous variables is 1, and 3 for the high setting.

(2) The amount of the correlation between the dependent variables, X_1, X_2, X_3, and X_4. The low setting is orthogonal, or zero correlation, while the high setting is 0.9, which is highly correlated.

(3) The variance of the extraneous variables. The low setting for the variance is 1, and the high setting is 100.

(4) The variance of the independent variables. The low setting for the variance is 1, and the high setting is 100.

(5) The sample size. The low setting for the sample size is deliberately near the threshold of the usefulness of the S_p criterion. A requirement of the S_p statistic is (n-p-2) should be greater than zero. At the low setting the sample

size is 10. The maximum value for p is 7, in the three extraneous variable case). When n equals 10 and p equals 7, the denominator of the S_D statistic is one. Therefore, the threshold for sample size is 10. If the sample size were any smaller, S_p could not be found for all possible subsets. Thus, the low setting for sample size is 10, while the high setting is 20.

(6) The variance of the ϵ term. The low setting for the variance of ϵ is 0.0625, and the high setting is 0.25. The computer program used to create the correlated data can be found in Appendix B.

When using Response Surface Methodology it is convenient to work with coded factors (-1, 1 variables) for the following reasons:

(1) By coding the factors, the resulting variables are of the same magnitude.

(2) The calculations necessary for the estimates are simplified.

(3) The resulting design matrix, **Z**, is orthogonal. As a result, the stepwise procedure can be used to find the significant factors with confidence.

The following equations are necessary to code the variables:

Z_1 = number of extraneous vars - 2

Z_2 = (correlation - 0.45) / 0.45

Z_3 = (variance of extraneous vars - 50.5) / 49.5

Z_4 = (variance of independent vars- 50.5) / 49.5

Z_5 = (sample size - 15) / 5

Z_6 = (variance of ϵ term - 0.15625) / 0.09375

where $Z_1,\ldots,Z_6$ are the coded variables.

Table I.
Variable Coding for Response Surface Methodology

Non-Coded			Coded		
Variable	Low	High	Variable	Low	High
Number of extraneous vars	1.0	3.0	Z_1	-1	1
Correlation	0.0	0.9	Z_2	-1	1
Variance of extraneous vars	1.0	100.0	Z_3	-1	1
Variance of independent vars	1.0	100.0	Z_4	-1	1
Sample Size	10.0	20.0	Z_5	-1	1
variance of ϵ term	0.0625	0.25	Z_6	-1	1

Experiment. It seems reasonable to assume significance in interactions between factors. For example, there might be a significant interaction between two of the main factors mentioned above. To insure the accurate calculation of estimates for both the main factors and

factor interactions, a full 2^6 factorial design is necessary. To construct the design matrix for a full factorial design, the main factors are varied from their low settings to their high settings in binary fashion. The interaction terms are simply the product of the corresponding main factors. An example of this process using a 2^2 full factorial design is summarized in the table below.

Table II.
Example of Coding Interaction Variables

Z_1	Z_2	Z_1Z_2 (interaction)
-1	-1	1
-1	1	-1
1	-1	-1
1	1	1

If a design with less than 2^6 runs is used, information on some of the interactions would be unobtainable.

Each of the 2^6 runs contains 60 replications. The raw data was obtained by running regressions on unprocessed data sets from Appendix B. The code necessary to perform the regression analysis is found in Appendix C. A FORTRAN program was written to accumulate the statistics. The program and a sample output can be found in Appendix D and E respectively. The following statistics were collected for each of the 64 runs.

(1) The average number of variables chosen using each technique.

(2) The average number of correct and extraneous variables of those chosen by the technique.

Results

Since the design matrix for this experiment is orthogonal, stepwise regression is used to select the significant factors. Resulting equations strictly screen factors which are most significant to their respective response. The equations should not be used for predicting the percentage of correct variables chosen. Several unsuccessful transformations were attempted to obtain a reasonable model, however none made common sense. Many models appeared to fit well using several common criteria (R^2, Adj R^2, residual plots, etc), however, the models were not logical. Using stepwise regression with a significance level of 0.01, yielded the equations given below. Again, the role of these equations is restricted to determining the most significant factors for a comparison of the three criteria.

Min MSE. For the Minimum MSE case, only two main factors were significant. They are the number of extraneous variables in the variable pool and the sample size. Since the variables are coded and the values of the standard errors for all the estimates are equal, only one value is printed for the standard errors.

$$R^2 = .9411$$
$$\text{Adj } R^2 = .9392$$

ymse(% correct) = .64448797 - .120997 (# ex)
+ .010415(sample)

standard errors = .0039

where

ex is the number of extraneous variables allowed in the variable pool.

sample is sample size of each replication for a given run.

This equation provides insight into the usefulness of the Min MSE criterion. Within the data region, the effect of increasing the sample size can be observed. For every "extra" data point, an increase of 1.04 percent of correct variables chosen is expected. This is useful information since sample size can be moderately controlled. The equation also indicates a decrease of 12 percent of correct variables chosen for every extraneous variable allowed into the variable pool. At first this information seems irrelevant since there is no way of telling <u>a priori</u> if a variable is relevant. However, it does emphasize the need to screen the variable pool before using the Min MSE criterion. It also shows the danger of the <u>kitchen sink</u> approach. The kitchen sink approach is where every possible

variable is tossed into the variable pool and the computer is allowed to make all the decisions. Therefore, the two significant main factors using the Min MSE criterion are sample size and the number of extraneous variables, and the number of extraneous variables is the most significant.

Min S_p. For the S_p case, two main factors and one three-way interaction are significant.

$$R^2 = .8811$$
$$\text{Adj } R^2 = .8751$$

$$\text{ysp(\% correct)} = .65598 - .1148(\#\text{ ex}) - .0184\ (\epsilon \text{ var}) - .0168(\text{int})$$

standard errors = .0056

where # ex is number of extraneous variables.

ϵ var is variance of the ϵ term.

int is the three-way interaction term.

The three terms are: number of extraneous variables, the ϵ variance and the variance of the X variables.

The information obtained here is acceptable but not quite as useful as in the MSE case. Apparently the most significant factors which effect S_p within the observed data space is the number of extraneous variables and the variation of Y. Again, the point to emphasize is as few extraneous variables as possible should be allowed in the variable pool. It is interesting to note that sample size

is insignificant at the α=.01 level. The low level set for sample size was near the threshold of S_p's usefulness on purpose. Consequently, S_p might be a good criterion to use when the sample size is small.

Min C_p. Unfortunately the resulting equation for the C_p case is not as descriptive as the two previous cases. The C_p equation included two main factors, one three way interaction, two four-way interactions, and a five-way interaction.

$$R^2 = .9280$$
$$\text{Adj } R^2 = .9202$$

$$\begin{aligned} ycp(\%correct) = {} & .65 - .11(\#ex) - .01\ (\epsilon var) \\ & -.01(\#ex, xvar, \epsilon var) \\ & +.02(cor, xvar, exvar, \epsilon var) \\ & + .01(\#ex, cor, xvar, exvar, \epsilon var) \end{aligned}$$

standard error = .0045

where

#ex	is the number of extraneous variables in the variable pool.
cor	is the amount of correlation between the independent variables.
exvar	is the variance of the extraneous variables.
xvar	is the variance of the independent variables.
ϵ var	is the variance of the ϵ term.

Only in this equation is the amount of correlation among the **X**'s significant. Even though it appears in the equation, it

is not one of the two main factors, but does appear in the four and five way interaction terms.

As is the case with the MSE equation and the S_p equation, the number of extraneous variables included in the variable pool is the most significant factor. However, probably the most notable result of this experiment is the repetition of variance factors appearing in the interactions. Actually this is expected, since the C_p statistic is heavily dependent on the total variation of the variable pool.

Comparison of Techniques

Up to this point, analysis has been limited to intra-technique. That is to say, the models constructed in the Response Surface phase show the effect of varying factors using a specific technique. Determining which technique actually outperforms the other two techniques is more useful.

Choice of a Comparison Criterion. Deciding which performance measure to use for comparing the Min MSE, Min C_p, and Min S_p variable selection criteria is not trivial. Due to the mechanics of the Min MSE procedure, the resulting subset of variables is always a super-set of both the Min C_p and Min S_p procedures. Also, in each of the sixty-four replications, the Min MSE procedure picked the highest percentage of correct variables of those chosen; additionally it picked the most correct variables in absolute terms. It seems the Min MSE criterion is superior

to the other two criterion. However, these statistics are a bit misleading.

In each of the sixty-four replications, the Min MSE criterion chose the most extraneous variables in absolute terms. Remembering the concept of best model versus true model, it is probably worse to have more extraneous variables in the model than the highest percentage of correct variables of those chosen. Therefore, comparing percentages to determine the "best technique" is not used.

Since there is no clear-cut criterion for comparing variable selection techniques, it was decided to compare how well each criterion performs using theoretical minimum mean square error of prediction (TMSEP).

$$TMSEP = \frac{\Sigma (Y_t - Y_p)^2}{(n - p)} \qquad (17)$$

where

Y_t is the theoretical value of the independent variable Y.

Y_p is the predicted value of Y, using one of the three criteria.

n is the sample size.

p is the number of parameters in the predicting equation.

There are two reasons for picking TMSEP as the performance measure. First, MSEP has received considerable attention during the past decade as the most promising criterion for variable selection. Second, the S_p criterion, which is based on MSEP, is praised and considered by many as the method which insures the model selected has the Min MSEP.

The TMSEP criterion is a variation of MSEP. Like MSEP, TMSEP calculates the squared difference between the predicted value and the actual value of the independent variable and adjusts the value for degrees of freedom. However, TMSEP differs from MSEP in its calculation. TMSEP is the squared difference between the actual value, from the underlying equation that is generating the independent variable (not including the error term, ϵ), and the predicted value using the model picked by the variable selection procedure. The resulting value is the theoretical mean square error of prediction (TMSEP).

At first, it appears this performance measure unfairly favors the Minimum C_p and Minimum S_p criteria. Both of these criteria are based on minimum MSEP. The difference between the two lies in the assumption that the regressors are fixed in the C_p case and are random in the S_p case. Since the regressors in this study are randomly generated, it is then assumed that the S_p criterion would outperform the other two. However, this is not necessarily true.

It is assumed when calculating the S_p and C_p statistics that all relevant variables are included in the variable pool. It is also assumed that variable pool does not contain extraneous variables. In this study both of these assumptions are violated. Therefore, it is possible that either the C_p or MSE criterion could outperform the S_p criterion. Since the MSE criterion is not designed to minimize MSEP, it is doubtful whether it would outperform the other two criteria using TMSEP as the performance measure.

Performance of the Selection Technique using TMSEP. Measurement of each criterion's performance is recorded under best-case and worst-case factor settings. If a technique outperformed the other two in both the best-case scenario and worst-case scenario, it is superior to the other two.

Since determining the necessary sample size is difficult, runs were made until the distributions for each technique stabilized. The number of necessary number of runs was 180 for both the best-case and worst-case scenarios.

Method of Comparison. For each of the 360 runs, the TMSEP value was calculated for every subset of variables. For example, when the factor setting for extraneous variables was one (best-case factor setting), there were four possible regressors. Since there were 2^k-1 subsets possible, in our example, fifteen TMSEP values were

calculated. In the three extraneous variable case the number of values increases to sixty-three. The 2^k-1 TMSEP values were then ordered from lowest to highest; lowest being more desirable. The TMSEP values are then calculated for the models chosen by each of the three techniques. Then, the values calculated from each technique's choice was compared to the rank ordered values and its ranking was recorded. For instance, if the value chosen by a technique was identical to the lowest value of TMSEPs, then a "one" was recorded. Additionally, the ranking of the TMSEP value for each technique was compared to the other two and recorded. A FORTRAN program was written to execute the procedure outlined in this section. The code for this program can be found in Appendix F, and a sample output can be found in Appendix G.

Best-Case Comparison. Using the equations found in the Response Surface phase, a common set of factors was obtained which constitutes a best-case scenario. The following settings maximize each techniques performance:

Table III.
Factor Settings For Best-Case Scenario

variable	Non-Coded value	coded value
Number of extraneous variables	1.0	-1.0
Correlation	0.9	+1.0
Variance of extraneous variables	100.0	+1.0
Variance of independent variables	1.0	-1.0
Sample size	20.0	+1.0
Variance of ϵ term	0.0625	-1.0

The Min S_p criterion was derived under the assumption that the regressors are random, while the Min C_p criterion was derived under the assumption that the regressors are fixed. Since the regressors are random in this experiment, it was expected that there would be a significant difference between the two. The following two figures summarize the results from the best-case comparison.

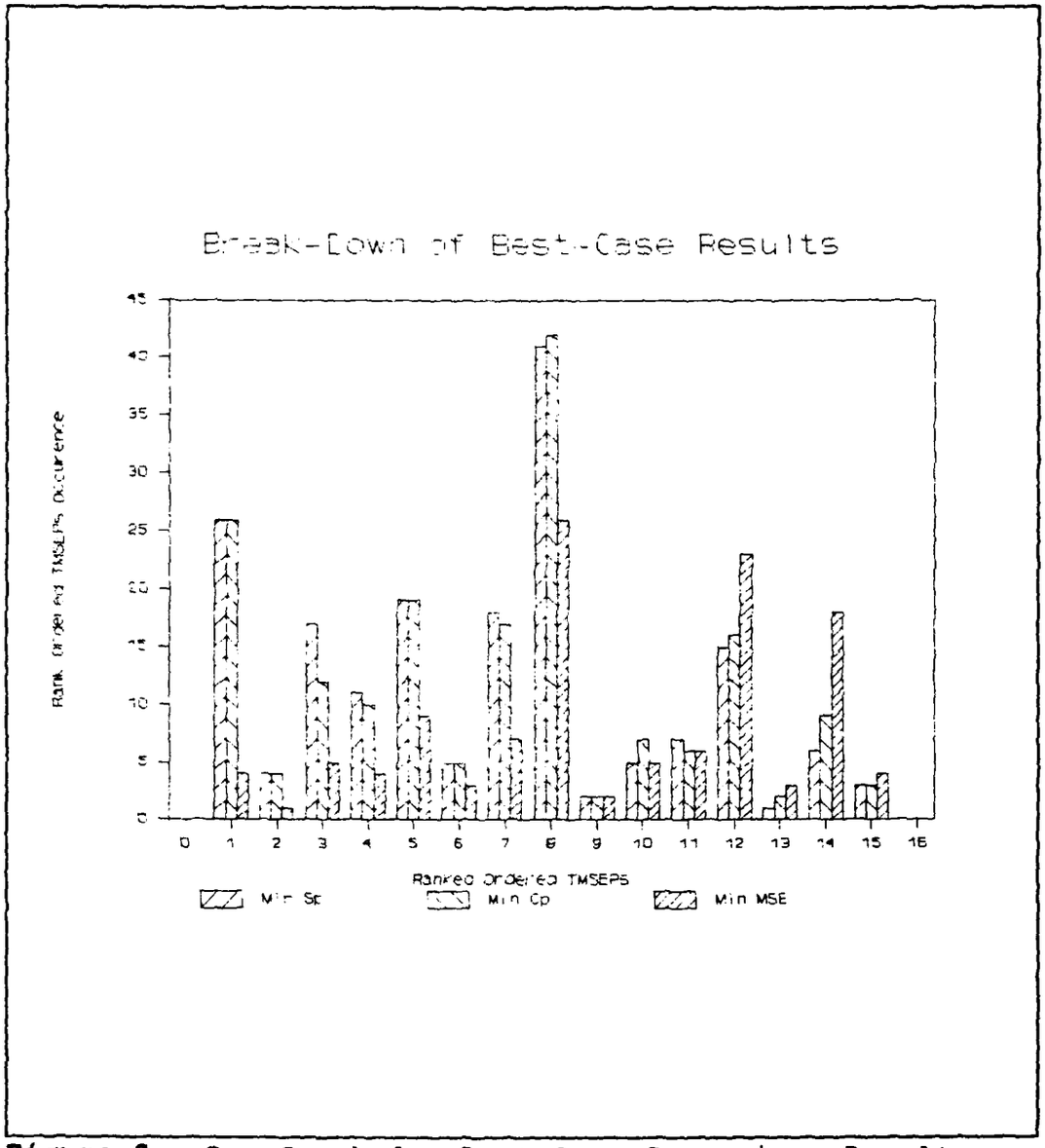

Figure 2. Bar Graph for Best-Case Comparison Results

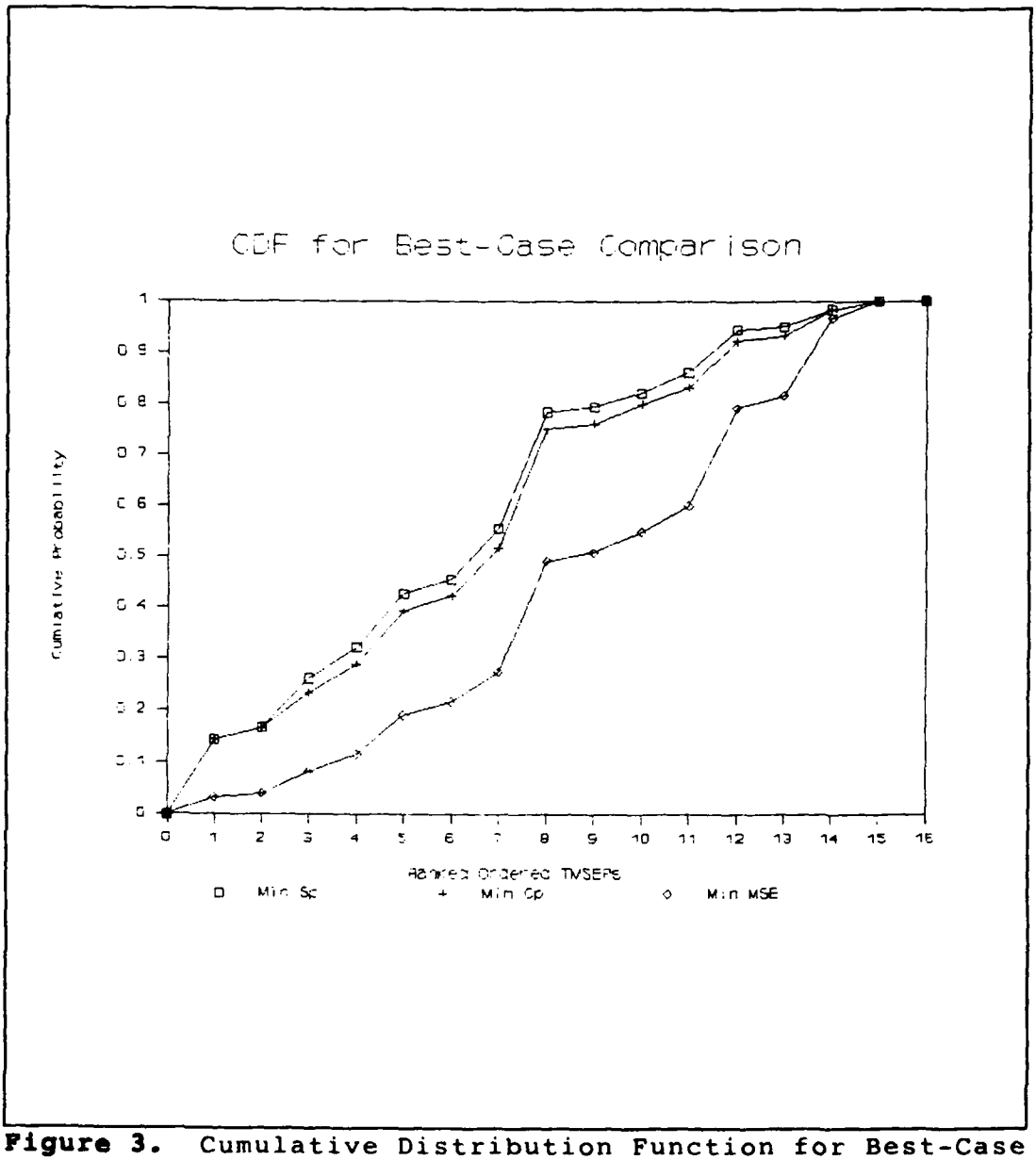

Figure 3. Cumulative Distribution Function for Best-Case Comparison

The cumulative distribution function of each technique indicates there probably is a significant difference between the two MSEP procedures, the difference between MSE and the other two is obvious. The chart below summarizes the number of times each technique chose the lowest TMSEP value amongst the three.

Table IV.
Best-Case Scenario Results

Criteria	Number of times Percentage Lowest Value	
C_p only	1	0.5%
S_p only	9	5.0%
MSE only	1	0.5%
C_p and S_p tied	58	32.0%
C_p, S_p, and MSE tied	111	62.0%

Allowing for ties, the following are the percentages that each technique found the lowest TMSEP value amongst the three.

Table V.
Best-Case Scenario Results

Criterion	Percentage of time lowest, or tied for lowest amongst the three
C_p	94.0%
S_p	99.0%
MSE	62.0%

The table above reaffirms our expectation that the Min S_p criterion is superior to the Min C_p criterion under a best-case scenario. To determine if the difference actually exists a hypothesis test is performed. The null hypothesis, Ho, is: the percentage of lowest TMSEP values chosen solely by the Min S_p criterion equals the percentage chosen solely by the Min C_p criterion. The alternative hypothesis, Ha, is: the percentage chosen solely by the Min S_p criterion is not equal to the percentage chosen solely by the Min C_p criterion.

$$Ho: Pcp = Psp$$
$$Ha: Pcp \neq Psp$$

$\alpha = 0.05$
critical probability = 0.039

The following is the equation used to calculate the critical probability:

$$2 \sum_{i=0}^{1} \binom{9}{i} \left(\frac{1}{2}\right)^{i} \left(1 - \frac{1}{2}\right)^{9-i}$$

where i is number of occurrences out of 9 trials, that C_p chose a model with a lower TMSEP value.

9 is the total number of times the two MSEP criteria differed.

This equation is based on the binomial distribution with a probability of 0.5. That is, the chance of one criterion picking the lowest TMSEP value is equal to the other.

Since the significance level is larger than the critical probability, the null hypothesis is rejected. The conclusion is there is a significant difference between the two criteria under the a best-case scenario, and the Min S_p is superior.

Worst-Case Comparison. By reversing the factor settings of the best-case scenario, a worst-case scenario is obtained.

Table VI.
Factor Settings For Worst-Case Scenario

variable	Non-Coded value	coded value
Number of extraneous variables	3.0	+1.0
Correlation	0.0	-1.0
Variance of extraneous variables	1.0	-1.0
Variance of independent variables	100.0	+1.0
Sample size	10.0	-1.0
Variance of ϵ term	0.25	+1.0

Similar to the best-case scenario results, the Min MSE criterion did significantly worse than the other two techniques. The graph below shows the cumulative distribution function for each of the criteria (the worst-case bar graphs are located in Appendix A).

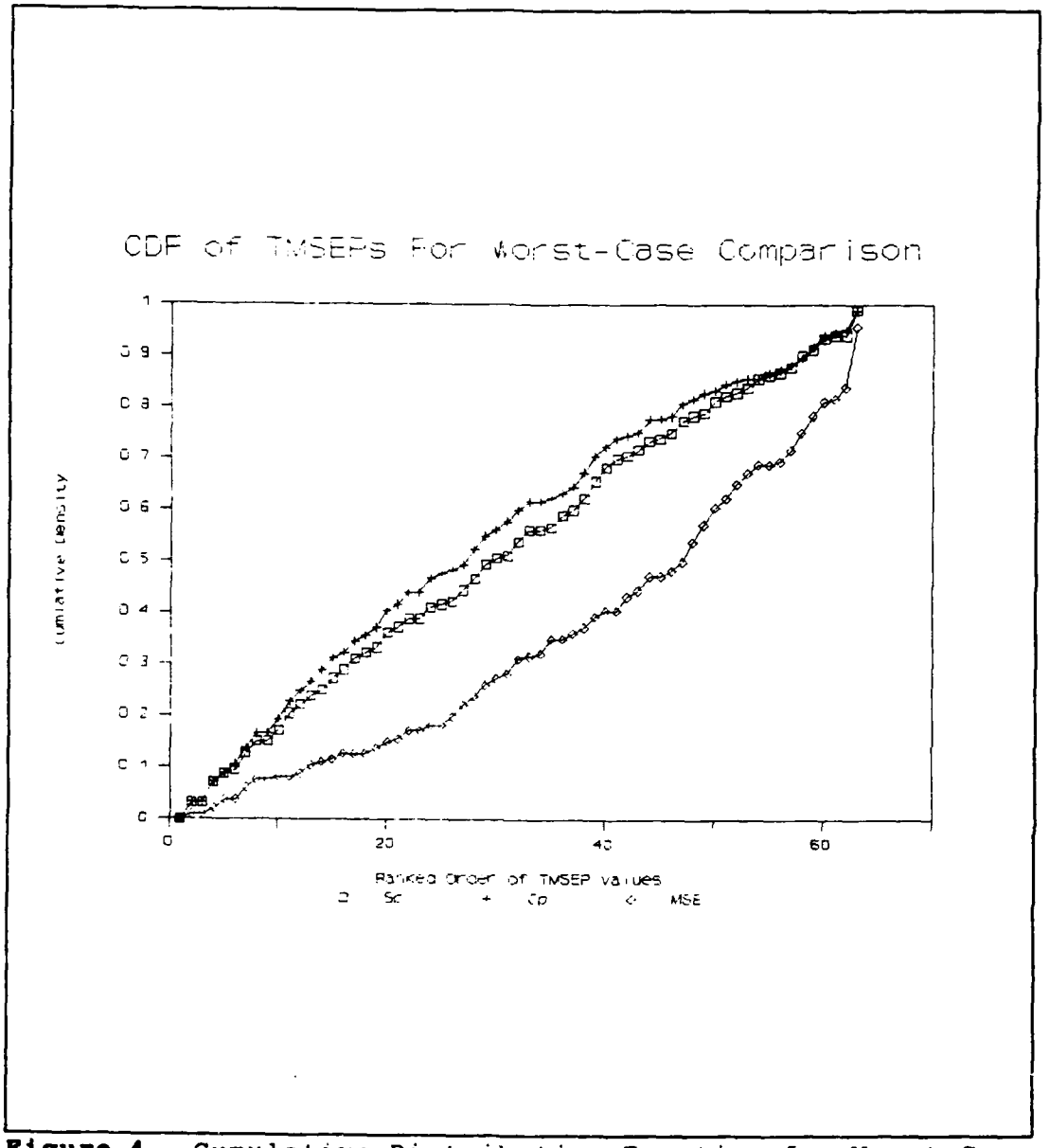

Figure 4. Cumulative Distribution Function for Worst-Case Comparison

Notice the cumulative distribution for C_p is larger than the other two criteria for most rank ordered TMSEP values. This is surprising since the factors are random. The question is whether the difference between the two MSEP criteria is significant. If so, the usefulness of the minimum S_p criterion would be questionable.

Table VII.
Worst-Case Scenario Results

Criteria	Number of times Lowest Value	Percentage
C_p only	28	16.0%
S_p only	4	2.0%
MSE only	0	0.0%
C_p and S_p tied	81	45.0%
C_p, S_p, and MSE tied	67	37.0%

Allowing for ties, the percentage each technique found the lowest TMSEP value amongst the three are summarized in the table below.

Table VIII.
Worst-Case Scenario Results

Criterion	Percentage of time lowest, or tied for lowest amongst the three
C_p	98.0%
S_p	91.0%
MSE	67.0%

Since the Min C_p criterion appears to be significantly better than the Min S_p criterion, a two tailed hypothesis test at the α=0.05 level was performed. This test was identical to the test in the best case scenario. The null hypothesis was: the percentage of lowest TMSEP values chosen solely by the Min S_p criterion is equal to the percentage chosen by the Min Cp criterion. While the alternative hypothesis was: the percentage chosen by the Min S_p criterion is not equal to the percentage chosen by the Min C_p criterion.

Ho: Pcp = Psp
Ha: Pcp ≠ Psp

α = 0.05
critical probability = 0.00002

The critical probability for this hypothesis test was calculated in exactly the same manner as in the best-case scenario. The equation necessary to find the critical probability is:

$$2 \sum_{i=0}^{4} \binom{32}{i} \left(\frac{1}{2}\right)^{i} \left(1 - \frac{1}{2}\right)^{32-i}$$

where

i is number of occurrences out of 32 trials, that C_p chose a model with a lower TMSEP value.

32 is the total number of times the two MSEP criteria differed.

Since the critical probability is less than the significance level, the null hypothesis is rejected. Thus, the conclusion is the Min C_p criterion significantly outperforms the Min S_p criterion in the worst-case scenario.

V. Conclusions and Further Research

The objectives of this research were: (1) identify some promising least squares selection procedures discussed in the literature during the previous decade, (2) use Response Surface Methodology to find the factors which most significantly drive the selection process for each of the these techniques, and (3) make a best-case and worst-case scenario comparison of the techniques.

Conclusion

The three techniques chosen, Min MSE, Min C_p, and Min S_p, have received praise in the last decade's literature. About ten years ago Min MSE was considered a favorable technique because of its similarity to the Max R^2 criterion, with an adjustment for degrees of freedom. In resent literature, the Min S_p and Min C_p criteria have received the majority of praise. Both techniques are based on the principle of minimizing the mean square errors of prediction. Of the two, the Min S_p criterion has received more attention. Its uses vary from finding the optimal number of variables to include in the model, to finding the best subset of variables. Unlike Min C_p, Min S_p is designed for random regressors, which is more practical.

During the Response Surface phase, emphasis was placed on the technique's performance under varying degrees of multicollinearity, variable variance, number of variables,

and sample size. A full 2^5 factorial design was constructed to find the most significant factors and their interactions. The three equations developed from the Response Surface phase show the relative weights of the most significant variables. These equations serve as a screening method, and should not be used to predict the percentage of correct variables of those chosen for any of the techniques.

The most noticeable result from the Response Surface phase dealt with the number of extraneous variables which were included in the variable pool. This factor was overwhelmingly the most significant in determining the percentage of correct variables chosen by a technique. In fact, if the number of extraneous variables was the only variable in any of the Response Surface equations, it would account for over 85% of the variation about the mean. Therefore, it is very important to screen the variables used in a regression analysis by some other means than just letting a computer program do the work.

The conclusion dealing with extraneous variables is consistent with Freedman (9), Flack and Chang (8), and Hoerl, Hoerl, Schuenemeyer (13), as well as Miller's suggested technique for a stepwise regression stopping-rule. It can not be over-emphasized that the "kitchen sink" approach is not the best way to go. One must decide which variables seem relevant to the dependent variable. If a

good reason for including a variable can not be determined, then it should not be included in the model.

Using common factors of the three equations found in the Response Surface phase, a best-case and worst-case scenario comparison was made using the theoretical mean square errors of prediction (TMSEP). Since the information necessary to find TMSEP is not available in practice, this statistic's usefulness is restricted to the simulation arena. However, if a technique such as (Min S_p) is suppose to minimize MSEP, then it is expected that it will also have one of the lowest TMSEPs of all possible models. Likewise, since Min C_p is also based on minimizing MSEP, it was expected to do fairly well.

The results of the best-case comparison indicate the Min S_p criterion performs better than Min C_p. Since the Min S_p criterion is by definition made for random regressors, one would expect it to do significantly better than Min Cp. Both criteria did very well in absolute terms. Over 75% of the TMSEP values calculated from the models chosen by S_p and C_p were in the upper half of rank ordered TMSEP values. 14% of the values found from the MSEP criteria's model were the absolute lowest TMSEP value possible. This is extremely good considering a fundamental regressor was dropped from the variable pool prior to using the variable selection criterion.

The results from the worst-case comparison were shocking. Surprisingly, Min C_p outperformed S_p. What makes this so surprising is the assumption that the regressors must be fixed to implement C_p. The difference between the C_p and S_p could be attributed to many things. First, when these two criteria were derived there was an assumption that all relevant regressors were included, and no extraneous variables were included in the variable pool. Both of these assumptions have been violated in this research. Under these circumstances, it is possible C_p outperforms S_p regardless of the nature of the regressors. In the best-case scenario, the MSE criterion performed significantly worse than the other two criteria.

The results of this thesis provide insight into the usefulness of the Min MSE, Min S_p, and Min C_p criteria under realistic conditions. It is possible that variables which significantly contribute to a model may be left out of the variable pool. It is also possible some extraneous variables will be included. The most important lesson of this research is: the number of extraneous variables which are in the variable pool significantly detriments the selection process. There must be some thought, commonly referred to by Operation Research analysts as the "art", used when picking variables to include in the variable pool. Not even the state-of-the-art variable selection criteria are able to perform well when the variables have not been

properly screened before implementation. This information may be useful to those who make a practice of collecting data on everything and letting the computer pick the "optimal model."

Another useful result of this research is the comparison of two MSEP criteria. A great deal of praise has been given to the minimum S_p criterion in the past decade. It has been identified as "one of the most promising" when the regressors are random, and one desires to minimize the mean square error of prediction. The minimum C_p criterion has also received praise for minimizing mean square error of prediction, but its usefulness is limited to cases where the regressors are fixed. Some have recommended that the Min C_p criterion should not be used in practice.

The results of this thesis indicate that the Min S_p criterion does, in fact, perform well when the circumstances are "nice." However, when circumstance are poor, the Min C_p criterion does better than S_p. Ironically, there is no way to be sure whether the conditions the variable selection techniques must perform under are best-case or worst-case, yet it seems like such information is necessary if one intends to minimize MSEP.

Recommendations for Further Research

A study of this intent lends itself to many follow-on studies. The methodology and groundwork are established but

embellishments will be necessary. Using more complex models would be one example. The model used in this study was simple: four regressors created the data with one eliminated from further consideration.

A full-blown simulation could be implemented to construct guidelines for usage of these criteria. For instance in this study, results indicate that if the circumstances are favorable, use the Min S_p criterion. However, if the circumstances are unfavorable, use the Min Cp criterion. By expanding the simulation to randomly generate many true variables and many extraneous variables and then eliminate a random number of variables, guidelines could be established for which technique should be used under various conditions that seem most likely. That is, if it is likely to be a situation where there are many extraneous variables and many variables omitted, use one technique, otherwise use another. Even though it is impossible to determine whether there are going to be many extraneous variables, one can assume if the dependent variable is something entirely new, like a flight characteristic of the Soviet's new Blackjack bomber, information on significant variables will be unobtainable. Likewise, some data collected probably will have little to do with the dependent variable. In such circumstances it would be useful to have a guideline for which variable selection criterion should be implemented. To completely examine the usefulness of each criterion and establish

guidelines, further research should include interaction terms and indicator variables.

One area which leads to further research deals with Response Surface. The data range used for this thesis was limited. It may be useful to expanding the Response Surface region to include negative correlation and larger sample sizes. If an indicator variable was used to contrast the effect of dropping a variable, the loop would be complete. That is, by including a variable to keep track of the difference between the full model and a model where a variable is dropped, one could quantify the effects of failing to collect data on all the significant variables. In this thesis, only information from dropping a variable was collected. It was assumed that if all variables were present, the techniques would perform better; plus similar simulation studies recorded results without dropping variables. However, it would be worthwhile to quantify the effects not including all significant variables into the variable pool.

This research has skimmed the surface of many myths associated with variable selection techniques, and in particular the usefulness of the C_p criterion. Some authors regard the Min C_p criterion as secondary to the Min S_p criterion. However, the results of this simulation do not support such ranking. The two criteria perform differently under certain circumstances. Under the best-case scenario,

the Min S_p criterion significantly outperform the Min C_p criterion. However, under the worst-case scenario, the Min C_p criterion significantly outperforms the S_p criterion.

Other simulations deal with the number of correct variables chosen of those available in absolute terms. No provisions are made for circumstances in which significant regressors are not included in the variable pool. Therefore, techniques praised as good variable section techniques may not be as appealing as originally thought. This research indicates this is the case with the Min S_p criterion.

Appendix A: Bar Graphs For Worst-Case Comparison

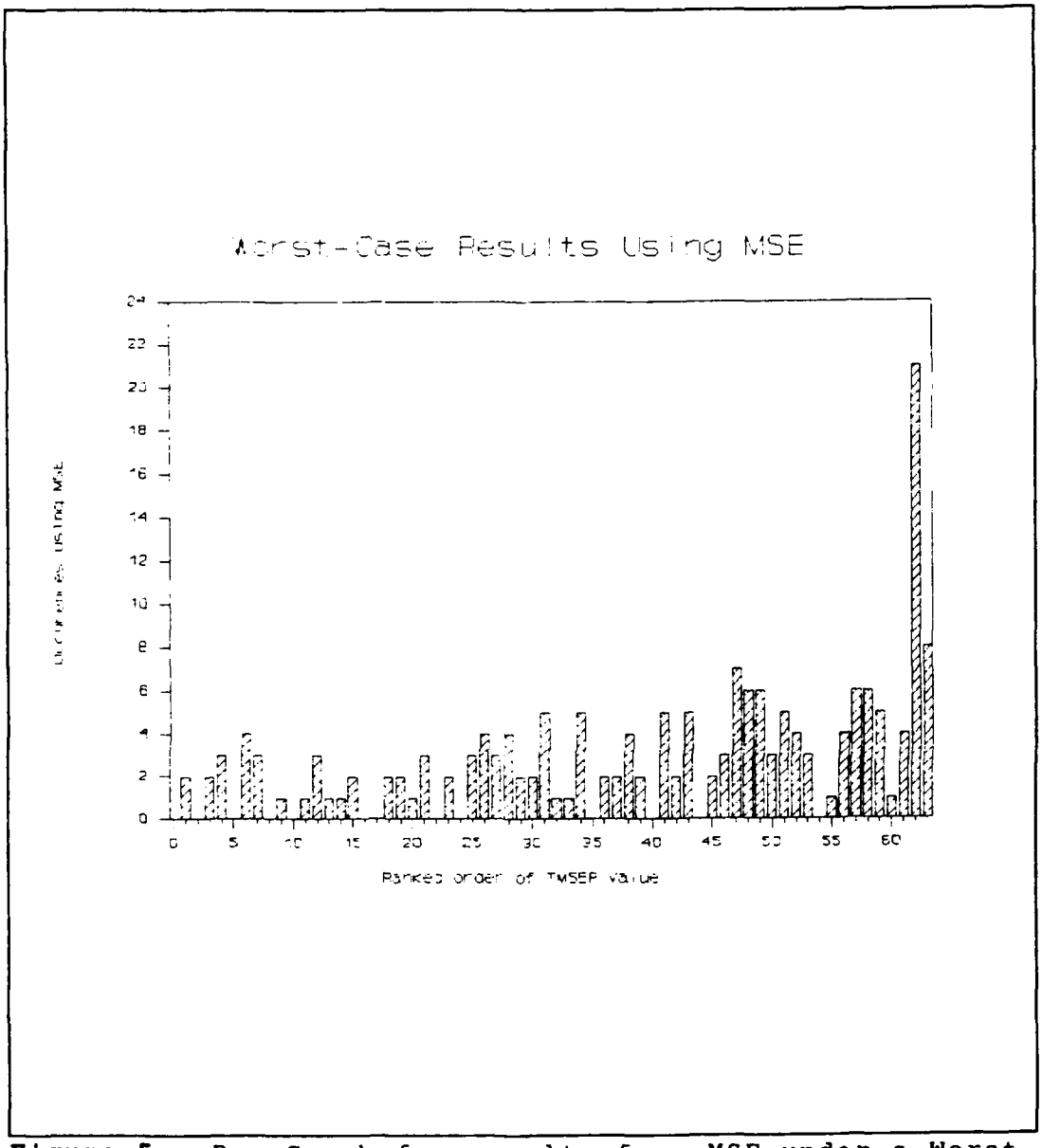

Figure 5. Bar Graph for results from MSE under a Worst-Case Scenario

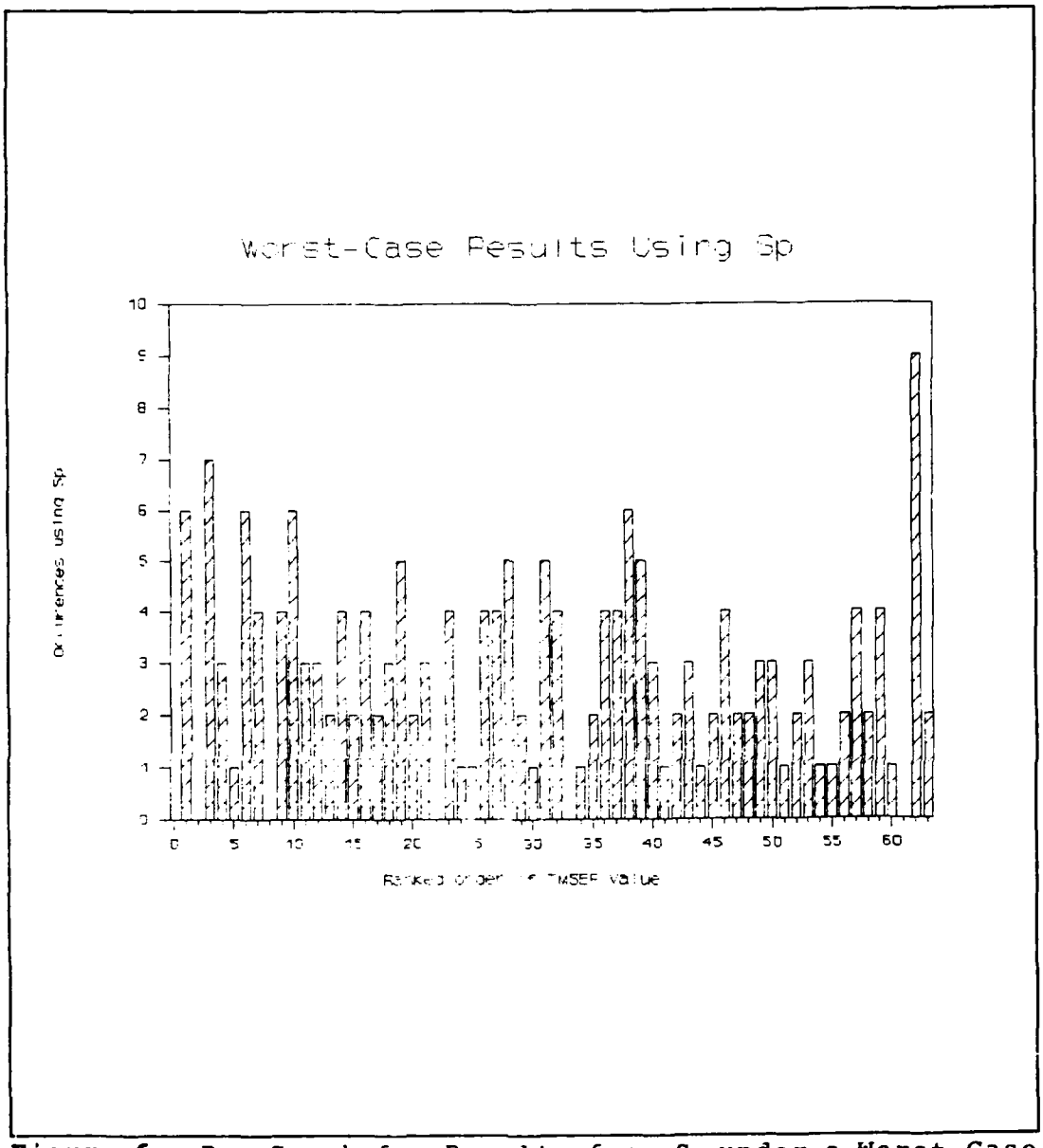

Figure 6. Bar Graph for Results from S_p under a Worst-Case Scenario

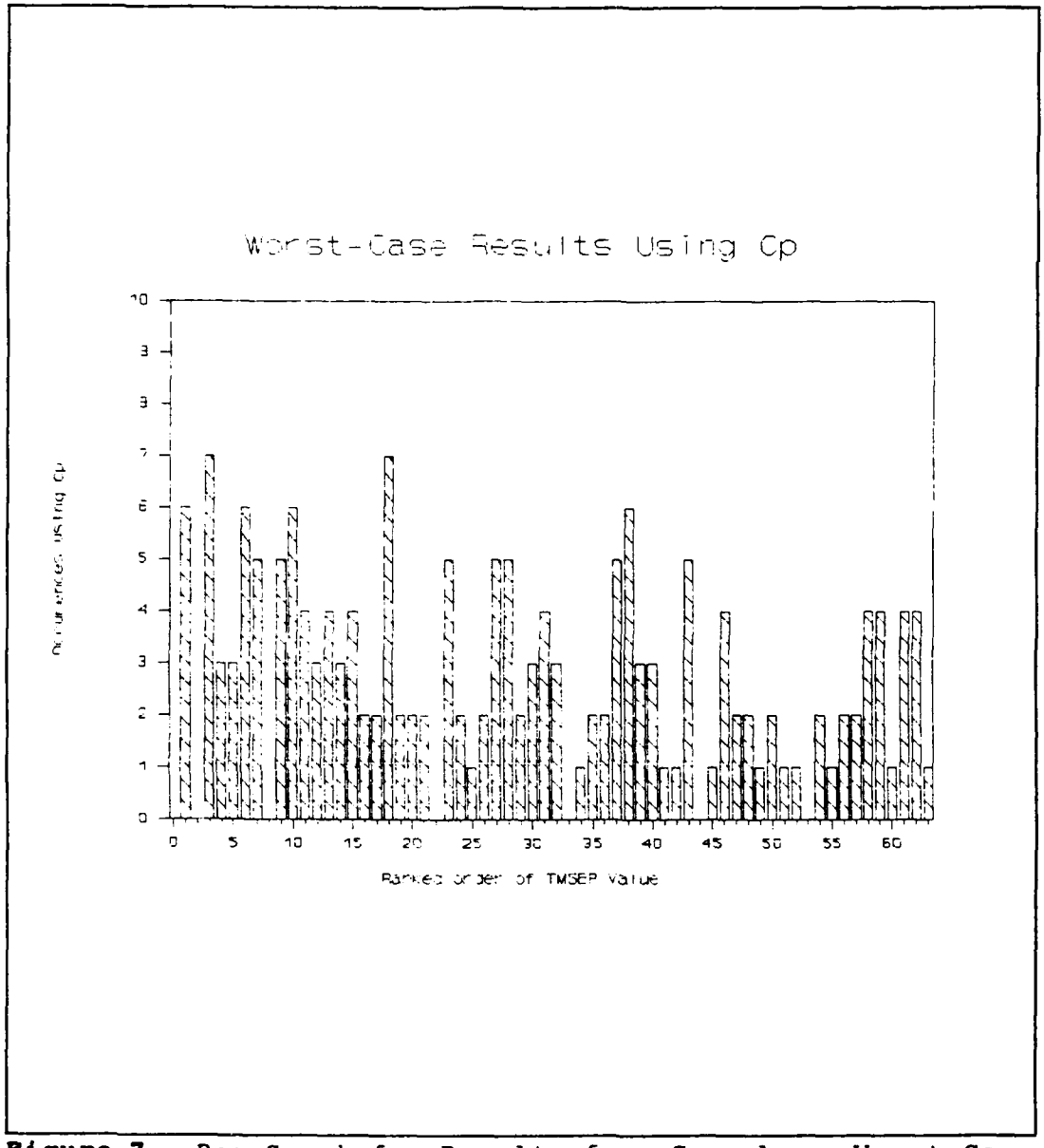

Figure 7. Bar Graph for Results from C_p under a Worst-Case Scenario

Appendix B: Macros to Generate Correlated Data Using S on the ASC or Success on the CSC

```
#############################################################
#
# This macro is designed to calculate correlated data from
# a random sample.  This macro is for the 1 extraneous
# variable case.  There is also a macro found for the 3
# extraneous variable case.
#
# To implement this program, the user must enter the S
# computer package by typing S on the ASC or Success on the
# CSC.
#
# Once in S, the following inputs must be made:
#
#    1. The 4x4 correlation matrix, K
#         > K_matrix(read(),4,4,byrow=T)
#         > "now input the rows of the matrix"
#
#    2. The standard deviation of the extraneous variables
#         > exerrstd_"value"
#
#    3. The standard deviation of the independent variables
#         > xerrstd_"value"
#
#    4. The standard deviation of the ∈ term
#         > yerrstd_"value"
#
#    5. The sample size (this is a two step procedure)
#         > ptsperset_"sample size"
#         > samplesize_"total number of points desired
#                       samplesize*number of runs
#
#    The macros can be implemented by the following commands
#         >source "make.1error"      for 1 ex vars case
#         >source "make.3error"      for 3 ex vars case
#
# The files which are implemented by these macros are the
#     following:
#
#       1. make.1error
#       2. make.3error
#       3. batch
#       4. four
#       5. matgen
#       6. qmake
# The data that is generated has an appended column of
#    integers which is used in the SAS runs.
#############################################################
```

```
#################################################################
#
# The following must be typed to write the data to a file.
#
#   > write(t(Z),"filename",#)
#      filename must be in quotes and # is the number of
#      columns of the data set.  In the 1ex case it is 7
#      in the 3ex case it is 9.
#
#################################################################

> # u an v are random noise, mean 0, and std exerrstd
> v_rnorm(samplesize,0,exerrstd)
> u_rnorm(samplesize,0,exerrstd)

> # x1..x4 are the independent variables

> x1_rnorm(samplesize,0,xerrstd)
> x2_rnorm(samplesize,0,xerrstd)
> x3_rnorm(samplesize,0,xerrstd)
> x4_rnorm(samplesize,0,xerrstd)
>
> X_cbind(x1,x2,x3,x4)
> e_eigen(K)
> lamb_diag(e$values)
> P_e$vectors
> XTX_t(X)%*X
> qq_e$values
> sig_diag(XTX)
> q1_qq/sig
> Q1_diag(sqrt(q1))
> Q_Q1%*t(P)
> Z_X%*Q
> op_rnorm(samplesize,0,yerrstd)
> y_Z[,1]+Z[,2]+Z[,3]+Z[,4]+op
> error_rnorm(samplesize,0,exerrstd)
> Z_cbind(y,Z,error)
>
>#  The following statements create the appended integers
>
> g_diag(diag(ptsperset)
> gg_c(g,g*2,g*3,g*4,g*5,g*6,g*7,g*8,g*9,g*10)
> gg_c(gg,gg+10,gg+20,gg+30,gg+40,gg+50)
> Z_cbind(gg,Z,v,u)
```

Appendix C: An All-Subsets SAS Program To Calculate all S_p, C_p, and MSE values

```
option linesize=80;
filename new 'msep.dat';
data new;
infile new;
input  set y x1 x2 x3 x4 e1;
proc rsquare data=new   mse sp cp b;
by set;
 model y= x1 x2 x3 e1  ;
```

Appendix D :Fortran Program to Find the Min MSE, Min C_p and Min S_p Values and Models

```
************************************************************
*  This program is designed to take a modified SAS listing
*  from Appendix C (the modification is simply deleting
*  all lines except for the raw data i.e. the number of
*  variables, the R2, the MSE, the Sp, and Cp values, as
*  well as the variables printed.  The program finds the
*  correct model using the Min MSE, Min Sp, and Min Cp
*  criteria.  In each case, the model chosen by each
*  technique is printed.  The values for each criteria are
*   also printed.
*
*  Statistics are collected on the average number of
*  variables chosen and the number of extraneous variables
*  chosen
*
*  variables num  is the number of variables in the model
*                    this is the first column of the SAS
*                    listing
*
*            i,j,k are counters for do loops.
*
*            ptrmse,ptrsp,ptrcp are variables used to keep
*                    track of the min values.
*
*            numcp,numsp,nummse are the average number
*                     of variables chosen by their
*                     respective technique.
*
*            cperr,sperr,mseeer are the average number
*                     of extraneous variables chosen
*                     by their respective technique.
*
*
*
************************************************************

      integer num(63),i,j,k,ptrmse,ptrsp
      integer ptrcp,varsmse,varssp,varscp
      integer check(6)
      integer n,emse,esp,ecp
      integer ccp,cmse,csp
      integer chartmse(0:3,0:3),chartcp(0:3,0:3)
      integer chartsp(0:3,0:3)
      real MSE(63),Sp(63),cp(63),r2(63)
      real minmse,minsp,mincp,nummse,numcp,numsp
```

```
      real mseeer,cpeer,speer
      character*2 m(6,63)
      check(1)=6
      check(2)=21
      check(3)=41
      check(4)=56
      check(5)=62
      check(6)=63

         do 7 i = 0,3
           do 3 k = 0,3
             chartmse(i,k)=0
             chartcp(i,k)=0
             chartsp(i,k)=0
3           continue
7           continue

      varsmse=0
      varssp =0
      varscp =0
      cumemse=0
      cumesp=0
      cumecp=0
      open (unit=10,file="temp",status='old')
      open (unit=11,file='OUT',status='new')

      do 20 k=1,60
      do 10 i=1,63
              emse=0
              esp=0
              ecp=0
              read(10,*)
num(i),r2(i),MSE(i),Sp(i),cp(i)
     +            ,(m(j,i),j=1,num(i))
              minmse=10000
              minsp =10000
              mincp =10000
              ptrmse=0
              ptrcp =0
              ptrsp =0
              do 30 j= 1,6
                if(mse(check(j)).lt.minmse) then
                  minmse=mse(check(j))
                  ptrmse=check(j)
                endif
                if(sp(check(j)).lt.minsp) then
                  minsp=sp(check(j))
                  ptrsp=check(j)
                endif
                if(cp(check(j)).lt.mincp) then
                  mincp=cp(check(j))
                  ptrcp=check(j)
```

```
                endif

30      continue
10        continue
            varsmse=varsmse+num(ptrmse)
            varssp =varssp +num(ptrsp)
            varscp =varscp +num(ptrcp)

          do 70 n=1,num(ptrmse)
            if(m(n,ptrmse).EQ.'E1') then
            emse=emse+1
            elseif(m(n,ptrmse).eq.'E2') then
            emse=emse+1
            elseif(m(n,ptrmse).eq.'E3') then
            emse=emse+1
            else
               continue
             endif

70         continue
          do 80 n=1,num(ptrsp)
            if(m(n,ptrsp).eq.'E1') then
            esp=esp+1
            elseif(m(n,ptrsp).eq.'E2') then
            esp=esp+1
            elseif(m(n,ptrsp).eq.'E3') then
            esp=esp+1
            else
            continue
            endif
80          continue
            do 90 n=1,num(ptrcp)
              if(m(n,ptrcp).eq.'E1') then
              ecp=ecp+1
              elseif(m(n,ptrcp).eq.'E2') then
              ecp=ecp+1
              elseif(m(n,ptrcp).eq.'E3') then
              ecp=ecp+1
              else
              continue
              endif
90            continue

             cumemse=cumemse+emse
             cumesp=cumesp+esp
             cumecp=cumecp+ecp
             cmse=num(ptrmse)-emse
             ccp=num(ptrcp)-ecp
             csp=num(ptrsp)-esp
             chartmse(cmse,emse)=chartmse(cmse,emse)+1
             chartcp(ccp,ecp)=chartcp(ccp,ecp)+1
```

```
            chartsp(csp,esp)=chartsp(csp,esp)+1



          write(11,*) 'MSE',num(ptrmse),mse(ptrmse)
     +      ,Sp(ptrmse),cp(ptrmse)
     +      ,(m(j,ptrmse),j=1,num(ptrmse))
           write(11,*) 'Sp',num(ptrsp),mse(ptrsp)
     +        ,Sp(ptrsp),cp(ptrsp),
     +        (m(j,ptrsp),j=1,num(ptrsp))
           write(11,*) 'Cp',num(ptrcp),mse(ptrcp)
     +        ,Sp(ptrcp),cp(ptrcp),
     +        (m(j,ptrcp),j=1,num(ptrcp))
        write(11,*) '****************************************'
        write(11,*) ' '
        write(11,*) ' '
20      continue
          nummse = real(varsmse)/60.0
          numsp  = real(varssp)/60.0
          numcp  = real(varscp)/60.0
          mseeer= real(cumemse)/60.0
          cpeer = real(cumecp) /60.0
          speer = real(cumesp) /60.0
         write(11,*) 'The avg number of vars using MSE
                       was ', nummse
         write(11,*) 'The avg number of errors from MSE
                       was', mseeer
         write(11,*) ' '
         write(11,*) 'The avg number of vars using Sp was',
                       numsp
         write(11,*) 'The avg number of errors from Sp was
                       ', speer
         write(11,*) '  '
         write(11,*) 'The avg number of vars using Cp was
                       ', numcp
         write(11,*) 'The avg number of errors from Cp was
                       ', cpeer
         write(11,*) '  '
         write(11,*) '  '
         write(11,*) ' MSE TABLE'
         write(11,*) '  '
             do 100 i=0,3
                 write(11,*) (chartmse(i,j),j=0,3)

100           continue
              write(11,*) ' '
              write(11,*) ' '
              write(11,*) 'Sp TABLE'
              write(11,*)
              do 110 i=0,3
                 write(11,*) (chartsp(i,j),j=0,3)
```

```
110             continue

                write(11,*) ' '
                write(11,*) ' '
                write(11,*) 'Cp TABLE'
                do 120 i=0,3
                  write(11,*) (chartcp(i,j),j=0,3)
120              continue
          END
```

Appendix E: Sample Output for Phase One of Analysis

The following is a sample of the output for one run (with 60 replications) for the Response Surface Methodology Phase (program found in Appendix D). The first column contains the technique used. The second column is the number of variables chosen. The third, fourth, and fifth columns are the values for Min MSE, Min C_p, and Min S_p respectively. The last column is the actual model chosen by the technique.

```
MSE   1  0.100654    1.43791e-02 -0.266093  X1
Sp    1  0.100654    1.43791e-02 -0.266093  X1
Cp    1  0.100654    1.43791e-02 -0.266093  X1
**************************************************


MSE   2    5.32424e-02    8.87374e-03    1.35403  X2  E1
Sp    1    6.00226e-02    8.57466e-03   0.898098  E1
Cp    1    6.00226e-02    8.57466e-03   0.898098  E1
**************************************************


MSE   3    7.63447e-02    1.52689e-02    3.17057  X1  X2  E1
Sp    1    7.64603e-02    1.09229e-02   0.904536  X2
Cp    1    7.64603e-02    1.09229e-02   0.904536  X2
**************************************************


MSE   1    2.95358e-02    4.21940e-03 -0.285088  X3
Sp    1    2.95358e-02    4.21940e-03 -0.285088  X3
Cp    1    2.95358e-02    4.21940e-03 -0.285088  X3
**************************************************


MSE   2    4.95487e-02    8.25812e-03    1.35192  X2  X3
Sp    2    4.95487e-02    8.25812e-03    1.35192  X2  X3
Cp    1    5.93601e-02    8.48001e-03    1.32763  X2
**************************************************


MSE   2    6.12631e-02    1.02105e-02    1.09398  X2  E1
Sp    1    6.46866e-02    9.24094e-03   0.147028  X2
Cp    1    6.46866e-02    9.24094e-03   0.147028  X2
```

```
**************************************************

MSE  2   3.19134e-02   5.31889e-03   1.45195  X2  X3
Sp   1   3.24903e-02   4.64148e-03  0.343450  X2
Cp   1   3.24903e-02   4.64148e-03  0.343450  X2
**************************************************

MSE  1   6.24057e-02   8.91510e-03 -0.299912  X3
Sp   1   6.24057e-02   8.91510e-03 -0.299912  X3
Cp   1   6.24057e-02   8.91510e-03 -0.299912  X3
**************************************************

MSE  1   3.94331e-02   5.63330e-03 -0.741483  X2
Sp   1   3.94331e-02   5.63330e-03 -0.741483  X2
Cp   1   3.94331e-02   5.63330e-03 -0.741483  X2
**************************************************

MSE  1   2.25172e-02   3.21674e-03   1.02995  E1
Sp   1   2.25172e-02   3.21674e-03   1.02995  E1
Cp   1   2.25172e-02   3.21674e-03   1.02995  E1
**************************************************

MSE  2   2.43904e-02   4.06507e-03   1.66245  X2  X3
Sp   1   2.54640e-02   3.63772e-03  0.756217  X2
Cp   1   2.54640e-02   3.63772e-03  0.756217  X2
**************************************************

MSE  2   3.30050e-02   5.50083e-03   1.29611  X1  X2
Sp   2   3.30050e-02   5.50083e-03   1.29611  X1  X2
Cp   2   3.30050e-02   5.50083e-03   1.29611  X1  X2
**************************************************

MSE  2   3.95879e-02   6.59799e-03   1.55024  X1  X2
Sp   2   3.95879e-02   6.59799e-03   1.55024  X1  X2
Cp   2   3.95879e-02   6.59799e-03   1.55024  X1  X2
**************************************************

MSE  1   8.97254e-02   1.28179e-02 -0.889888  X1
Sp   1   8.97254e-02   1.28179e-02 -0.889888  X1
Cp   1   8.97254e-02   1.28179e-02 -0.889888  X1
**************************************************
MSE  1   3.95394e-02   5.64849e-03 -0.658592  X2
Sp   1   3.95394e-02   5.64849e-03 -0.658592  X2
Cp   1   3.95394e-02   5.64849e-03 -0.658592  X2
```

```
****************************************************
MSE  2   3.60757e-02   6.01262e-03    1.61985  X2  E1
Sp   1   3.66450e-02   5.23500e-03   0.524038  E1
Cp   1   3.66450e-02   5.23500e-03   0.524038  E1
****************************************************

MSE  2   5.85844e-02   9.76407e-03    2.28246  X1  X2
Sp   2   5.85844e-02   9.76407e-03    2.28246  X1  X2
Cp   2   5.85844e-02   9.76407e-03    2.28246  X1  X2
****************************************************

MSE  2   4.41387e-02   7.35645e-03    1.80789  X2  X3
Sp   1   4.42947e-02   6.32782e-03   0.661062  X2
Cp   1   4.42947e-02   6.32782e-03   0.661062  X2
****************************************************

MSE  1  0.100894    1.44134e-02 -0.614306  X3
Sp   1  0.100894    1.44134e-02 -0.614306  X3
Cp   1  0.100894    1.44134e-02 -0.614306  X3
****************************************************
MSE  1   3.93769e-02   5.62527e-03  -0.790119  X3
Sp   1   3.93769e-02   5.62527e-03  -0.790119  X3
Cp   1   3.93769e-02   5.62527e-03  -0.790119  X3
****************************************************

MSE  1   4.66999e-02   6.67142e-03  -0.101598  E1
Sp   1   4.66999e-02   6.67142e-03  -0.101598  E1
Cp   1   4.66999e-02   6.67142e-03  -0.101598  E1
****************************************************

MSE  1   4.52970e-02   6.47101e-03  -0.946171  X2
Sp   1   4.52970e-02   6.47101e-03  -0.946171  X2
Cp   1   4.52970e-02   6.47101e-03  -0.946171  X2
****************************************************

MSE  3   2.84832e-02   5.69664e-03    3.30669  X1  X2  X3
Sp   1   3.08908e-02   4.41297e-03    1.67367  X2
Cp   1   3.08908e-02   4.41297e-03    1.67367  X2
****************************************************

MSE  1   4.47498e-02   6.39283e-03  -0.492574  E1
Sp   1   4.47498e-02   6.39283e-03  -0.492574  E1
Cp   1   4.47498e-02   6.39283e-03  -0.492574  E1
****************************************************
```

```
MSE  1   6.10468e-02   8.72097e-03 -0.754845  X2
Sp   1   6.10468e-02   8.72097e-03 -0.754845  X2
Cp   1   6.10468e-02   8.72097e-03 -0.754845  X2
*****************************************************


MSE  1   2.40736e-02   3.43908e-03   1.06484  X3
Sp   1   2.40736e-02   3.43908e-03   1.06484  X3
Cp   1   2.40736e-02   3.43908e-03   1.06484  X3
*****************************************************


MSE  1   4.34820e-02   6.21172e-03  0.251277  X1
Sp   1   4.34820e-02   6.21172e-03  0.251277  X1
Cp   1   4.34820e-02   6.21172e-03  0.251277  X1
*****************************************************


MSE  2   2.60387e-02   4.33979e-03   1.65346  X2  X3
Sp   2   2.60387e-02   4.33979e-03   1.65346  X2  X3
Cp   2   2.60387e-02   4.33979e-03   1.65346  X2  X3
*****************************************************


MSE  3   1.14088e-02   2.28177e-03   3.04763  X1  X3  E1
Sp   3   1.14088e-02   2.28177e-03   3.04763  X1  X3  E1
Cp   3   1.14088e-02   2.28177e-03   3.04763  X1  X3  E1
*****************************************************


MSE  4   1.11825e-02   2.79562e-03   5.00000  X1  X2  X3  E1
Sp   2   1.66321e-02   2.77202e-03   6.41136  X1  E1
Cp   4   1.11825e-02   2.79562e-03   5.00000  X1  X2  X3  E1

*****************************************************


MSE  3   1.24700e-02   2.49400e-03   3.16464  X2  X3  E1
Sp   1   1.44043e-02   2.05776e-03   1.95435  X2
Cp   1   1.44043e-02   2.05776e-03   1.95435  X2
*****************************************************
MSE  1   3.17437e-02   4.53481e-03 -0.727781  X3
Sp   1   3.17437e-02   4.53481e-03 -0.727781  X3
Cp   1   3.17437e-02   4.53481e-03 -0.727781  X3
*****************************************************


MSE  3   3.45629e-02   6.91257e-03   3.52295  X1  X2  X3
Sp   1   4.07384e-02   5.81977e-03   2.67969  X3
Cp   1   4.07384e-02   5.81977e-03   2.67969  X3
*****************************************************
```

```
MSE  1   6.70823e-02   9.58318e-03 -0.670246  X3
Sp   1   6.70823e-02   9.58318e-03 -0.670246  X3
Cp   1   6.70823e-02   9.58318e-03 -0.670246  X3
*********************************************************


MSE  2   3.87667e-02   6.46111e-03   1.12307  X3  E1
Sp   1   4.21533e-02   6.02190e-03  0.366424  E1
Cp   1   4.21533e-02   6.02190e-03  0.366424  E1
*********************************************************


MSE  1   4.98360e-02   7.11942e-03 -0.172478  X2
Sp   1   4.98360e-02   7.11942e-03 -0.172478  X2
Cp   1   4.98360e-02   7.11942e-03 -0.172478  X2
*********************************************************


MSE  1   1.98619e-02   2.83742e-03 -0.108219  E1
Sp   1   1.98619e-02   2.83742e-03 -0.108219  E1
Cp   1   1.98619e-02   2.83742e-03 -0.108219  E1
*********************************************************


MSE  2   2.27566e-02   3.79276e-03   1.19987  X1  X2
Sp   2   2.27566e-02   3.79276e-03   1.19987  X1  X2
Cp   1   2.67326e-02   3.81894e-03  0.981005  X2
*********************************************************


MSE  1   2.31557e-02   3.30796e-03 -0.325469  X2
Sp   1   2.31557e-02   3.30796e-03 -0.325469  X2
Cp   1   2.31557e-02   3.30796e-03 -0.325469  X2
*********************************************************


MSE  2   2.96808e-02   4.94681e-03   1.36022  X1  X3
Sp   2   2.96808e-02   4.94681e-03   1.36022  X1  X3
Cp   2   2.96808e-02   4.94681e-03   1.36022  X1  X3
*********************************************************


MSE  3   2.72695e-02   5.45390e-03   3.00697  X1  X2  E1
Sp   3   2.72695e-02   5.45390e-03   3.00697  X1  X2  E1
Cp   3   2.72695e-02   5.45390e-03   3.00697  X1  X2  E1
*********************************************************


MSE  2   4.81207e-02   8.02012e-03   1.77244  X3  E1
Sp   1   5.14718e-02   7.35311e-03   1.05649  X3
Cp   1   5.14718e-02   7.35311e-03   1.05649  X3
```

```
*****************************************************

MSE  1  0.124977    1.78539e-02 -0.908955  X2
Sp   1  0.124977    1.78539e-02 -0.908955  X2
Cp   1  0.124977    1.78539e-02 -0.908955  X2
*****************************************************

MSE  1    3.86961e-02    5.52802e-03 -0.554765  X3
Sp   1    3.86961e-02    5.52802e-03 -0.554765  X3
Cp   1    3.86961e-02    5.52802e-03 -0.554765  X3
*****************************************************

MSE  2    1.07762e-02    1.79603e-03   1.05727  X3  E1
Sp   1    1.25208e-02    1.78869e-03  0.715458  E1
Cp   1    1.25208e-02    1.78869e-03  0.715458  E1
*****************************************************

MSE  4    2.31882e-02    5.79705e-03   5.00000  X1  X2  X3
E1
Sp   2    2.46525e-02    4.10876e-03   3.44206  X1  E1
Cp   2    2.46525e-02    4.10876e-03   3.44206  X1  E1
*****************************************************

MSE  1    6.04122e-02    8.63031e-03 -0.792417  X1
Sp   1    6.04122e-02    8.63031e-03 -0.792417  X1
Cp   1    6.04122e-02    8.63031e-03 -0.792417  X1
*****************************************************

MSE  1    4.62591e-02    6.60844e-03  0.381250  E1
Sp   1    4.62591e-02    6.60844e-03  0.381250  E1
Cp   1    4.62591e-02    6.60844e-03  0.381250  E1
*****************************************************

MSE  1  0.110044    1.57205e-02 -0.301020  X1
Sp   1  0.110044    1.57205e-02 -0.301020  X1
Cp   1  0.110044    1.57205e-02 -0.301020  X1
*****************************************************
MSE  1  0.116894    1.66991e-02 -0.512063  E1
Sp   1  0.116894    1.66991e-02 -0.512063  E1
Cp   1  0.116894    1.66991e-02 -0.512063  E1
*****************************************************
MSE  3    5.91980e-02    1.18396e-02    3.09286  X1  X3  E1
Sp   1    6.13972e-02    8.77102e-03    1.04274  X2
Cp   1    6.13972e-02    8.77102e-03    1.04274  X2
*****************************************************
```

```
MSE  3   1.67376e-02   3.34753e-03    3.37179  X2  X3  E1
Sp   3   1.67376e-02   3.34753e-03    3.37179  X2  X3  E1
Cp   3   1.67376e-02   3.34753e-03    3.37179  X2  X3  E1
*****************************************************

MSE  1   8.60897e-02   1.22985e-02   0.261788  E1
Sp   1   8.60897e-02   1.22985e-02   0.261788  E1
Cp   1   8.60897e-02   1.22985e-02   0.261788  E1
*****************************************************

MSE  3   1.19855e-02   2.39709e-03    3.05798  X1  X2  X3
Sp   2   1.26033e-02   2.10055e-03    2.20518  X1  X3
Cp   2   1.26033e-02   2.10055e-03    2.20518  X1  X3
*****************************************************

MSE  2   3.42009e-02   5.70015e-03    1.36503  X1  E1
Sp   1   3.86254e-02   5.51791e-03   0.924681  E1
Cp   1   3.86254e-02   5.51791e-03   0.924681  E1
*****************************************************

MSE  1   3.65205e-02   5.21721e-03  -9.18206e-02  X3
Sp   1   3.65205e-02   5.21721e-03  -9.18206e-02  X3
Cp   1   3.65205e-02   5.21721e-03  -9.18206e-02  X3
*****************************************************

MSE  2   2.50195e-02   4.16992e-03    1.49327  X2  E1
Sp   1   2.60450e-02   3.72071e-03   0.535332  E1
Cp   1   2.60450e-02   3.72071e-03   0.535332  E1
*****************************************************

MSE  1   4.73777e-02   6.76825e-03  -0.485272  E1
Sp   1   4.73777e-02   6.76825e-03  -0.485272  E1
Cp   1   4.73777e-02   6.76825e-03  -0.485272  E1
*****************************************************
MSE  2   4.89740e-02   8.16234e-03    1.74329  X1  X3
Sp   1   5.48795e-02   7.83993e-03    1.35524  X1
Cp   1   5.48795e-02   7.83993e-03    1.35524  X1
*****************************************************

MSE  1   2.25924e-02   3.22748e-03   0.729942  E1
Sp   1   2.25924e-02   3.22748e-03   0.729942  E1
Cp   1   2.25924e-02   3.22748e-03   0.729942  E1
*****************************************************
```

```
The avg number of vars using MSE was     1.71667
The avg number of errors from MSE was   0.416667

The avg number of vars using Sp was      1.26667
The avg number of errors from Sp was    0.333333

The avg number of vars using Cp was      1.26667
The avg number of errors from Cp was    0.333333
```

where

errors stands for extraneous terms

vars stands for the number of variables chosen.

To find the response in the Response Surface Phase, the following equation was used.

(1-avg errors)/(avg vars)

For example, the following is the calculation for the C_p case.

(1-.33333)/1.2667 = .5263

This value was used as one of the 64 C_p response values in the response surface phase.

Appendix F: Fortran Program for finding TMSEP values, rank ordering them and Comparing the three criteria

```
************************************************************
*
*
* This program is designed to take a modified SAS program
* and an existing data set and find the TMSEP the models.
* The modifications necessary are the same as discussed in
* Appendix D.
*
* The one exception is that the SAS listing prints out the
* coefficient values.  When the variable is not included in
* the model a "." is left.  This symbol must be replaced
* with a "0".
* To accomplish this the following UNIX editing command can
* be used.
*
*   g/ \. /s// 0 /g
*
*  This command will make the appropriate substitution.
*
*  In this program all integer values are counters (for do
*  loops except for the arrays.
*
*  variable  ptrmse,ptrsp,ptrcp are used to keep track
*                  of the lowest value found using
*                  MSE, Sp, and Cp respectively.
*
*             spcount,equalcount,spcpct,cpcount are all
*                    counters for keeping track of the
*                    number of times the lowest TMSEP value
*                    was found by Min Sp, all techniques,
*                    Min Sp and Min Cp, and Min Cp only.
*
*             r2,mse,sp,cp are the arrays for r2, mse, Sp
*                    and Cp.
*
*             b0 and betas are the values for the
*                    coefficients
*
*             yssepmse,yssepsp,yssepcp are the Sum-of
*                    -Squares for the respective technique.
*
*             ymsepmse,ymsepsp,ymsepcp are the MSEP
*                 values using the respective technique.
*
*
************************************************************

      integer i,k,j,p,h,num(15),ptrmse,ptrsp,ptrcp
```

```
      integer spcount,equalcnt,set,check(4),spcpct,cpcount
      real b0(15),r2(15),mse(15),sp(15)
      real betas(4,15),minmse,mincp,minsp
      real x(4),ex,x3ex3(4),b0mse
      real b0sp,b0cp,ypredmse,ypredsp
      real ypredcp,ymsepmse,ymsepsp,ymsepcp
      real yssepmse,yssepsp,yssepcp,y
      real cp(15)

      spcpct=0
      equalcnt=0
      spcount=0
      cpcount=0
      msecount=0
      check(1)=4
      check(2)=10
      check(3)=14
      check(4)=15

      open(unit=12,file= 'msep1.lis',status='old')
      open(unit=10,file='msep1.dat',status='old')
      open(unit=11,file='msep1.out',status='new')

      do 20 k=1,60
      do 10 i=1,15
      read (12,*)  num(i), r2(i), mse(i), sp(i), cp(i),
b0(i)
     + ,betas(1,i),betas(2,i),betas(3,i),betas(4,i)
        minmse = 10000
        minsp  = 10000
        mincp  = 10000
        yssepmse=0
        yssepsp =0
        yssepcp =0
        ptrmse = 0
        ptrsp  = 0
        ptrcp  = 0
        do 30 j=1,4
           if(mse(check(j)).lt.minmse) then
              minmse = mse(check(j))
              ptrmse = check(j)
           endif
           if (sp(check(j)).lt.minsp) then
              minsp  = sp(check(j))
              ptrsp  = check(j)
           endif
           if (cp(check(j)).lt.mincp) then
              mincp = cp(check(j))
              ptrcp = check(j)
           endif
30       continue
10     continue
```

```
      do 50 h= 1,20
        read (10,*) set,y,x(1),x(2),x(3),x(4),ex
        ypredmse= 0
        ypredsp = 0
        ypredcp = 0
        yactual = 0
        b0mse   = b0(ptrmse)
        b0sp    = b0(ptrsp)
        b0cp    = b0(ptrcp)
        ypredmse= b0mse
        ypredsp = b0sp
        ypredcp = b0cp
        x3ex3(1)= x(1)
        x3ex3(2)= x(2)
        x3ex3(3)= x(3)
        x3ex3(4)= ex

        do 60 p=1,4
           yactual = yactual+x(p)
60      continue
        do 70 p=1,4
           ypredmse = ypredmse + betas(p,ptrmse)*x3ex3(p)
           ypredsp  = ypredsp  + betas(p,ptrsp) *x3ex3(p)
           ypredcp  = ypredcp  + betas(p,ptrcp) *x3ex3(p)
70      continue
        yssepmse = ((ypredmse-yactual)**real(2)) + yssepmse

        yssepsp  = ((ypredsp -yactual)**real(2)) + yssepsp
        yssepcp  = ((ypredcp -yactual)**real(2)) + yssepcp
50    continue
        ymsepmse  = yssepmse / (20-num(ptrmse))
        ymsepsp   = yssepsp  / (20-num(ptrsp))
        ymsepcp   = yssepcp  / (20-num(ptrcp))

        if(ymsepsp.lt.ymsepcp) spcount=spcount+1
        if(ymsepsp.eq.ymsepmse) equalcnt=equalcnt+1
        if(ymsepcp.lt.ymsepsp) cpcount=cpcount+1
        if(ymsepmse.lt.ymsepsp) msecount=msecount+1
        if(ymsepsp.eq.ymsepcp.and.ymsepmse.gt.ymsepsp)
      spcpcnt=spcpct+1
        write(11,*) 'run        ',k
        write(11,*) ' '
        write(11,*) ' '
        write(11,*) ' Sp  MSEP  =  ', ymsepsp
        write(11,*) ' Cp  MSEP  =  ', ymsepcp
        write(11,*) ' MSE MSEP =   ', ymsepmse
        write(11,*) ' '
        write(11,*) ' '
20    continue
      write(11,*) ' '
      write(11,*) ' '
      write(11,*) ' '
```

```
write(11,*) ' Number of times Sp  had smallest MSEP
                ',spcount
write(11,*) ' Number of times Cp  had smallest MSEP
                ',cpcount
write(11,*) ' Number of times MSE had smallest MSEP
                ',msecount
write(11,*) ' Number of times Cp and Sp were equal
                ',spcpct
write(11,*) ' All MSEPs were equal',equalcnt
END
```

Appendix G: Sample Output for Phase Two of Analysis

The following is a sample output for phase two of the analysis. This output is from the program found in Appendix F. The TMSEP values from each method is given, then the rank ordered TMSEP values for each subset of values is given.

```
run                    1

 Sp  TMSEP  =      8.1625051E-04
 Cp  TMSEP  =      8.1625051E-04
 MSE TMSEP  =      8.1625051E-04

TMSEPs FOR RUN          1

          1  5.3875311E-04
          2  6.2458171E-04
          3  6.2816928E-04
          4  7.8544696E-04
          5  8.1625051E-04
          6  8.6704991E-04
          7  8.7211549E-04
          8  8.9747010E-04
          9  9.1317965E-04
         10  9.5657917E-04
         11  1.3182798E-03
         12  1.4625470E-03
         13  1.7037858E-03
         14  1.8694992E-03
         15  9.8459609E-03
run                    2

 Sp  TMSEP  =      3.9654067E-03
 Cp  TMSEP  =      3.9654067E-03
 MSE TMSEP  =      3.9654067E-03

TMSEPs FOR RUN          2
```

```
            1   2.0111492E-03
            2   2.2425181E-03
            3   2.2925893E-03
            4   3.9654067E-03
            5   4.0591503E-03
            6   4.1853175E-03
            7   4.3666265E-03
            8   4.4225999E-03
            9   4.4807326E-03
           10   4.4842619E-03
           11   5.8024856E-03
           12   5.8971639E-03
           13   6.2541468E-03
           14   6.5273000E-03
           15   1.0629051E-02
run                      3

 Sp  TMSEP  =       1.7187677E-02
 Cp  TMSEP  =       1.7187677E-02
 MSE TMSEP  =       1.7187677E-02


TMSEPs FOR RUN              3

            1   1.0465025E-02
            2   1.2726073E-02
            3   1.3100259E-02
            4   1.6458001E-02
            5   1.7187677E-02
            6   1.7703351E-02
            7   1.9149231E-02
            8   2.0130737E-02
            9   2.1833103E-02
           10   2.2608828E-02
           11   2.3035588E-02
           12   2.3823561E-02
           13   2.5661029E-02
           14   2.7239855E-02
           15   3.0690275E-02
run                      4

 Sp  TMSEP  =       2.8920151E-02
 Cp  TMSEP  =       2.8920151E-02
 MSE TMSEP  =       2.8920151E-02
```

TMSEPs FOR RUN 4

1	1.8989012E-02
2	1.9804025E-02
3	2.0740228E-02
4	2.1199832E-02
5	2.1934517E-02
6	2.2168593E-02
7	2.2491228E-02
8	2.3455916E-02
9	2.3542039E-02
10	2.4646813E-02
11	2.5090951E-02
12	2.6086060E-02
13	2.6278302E-02
14	2.8023606E-02
15	2.8920151E-02

run 5

Sp TMSEP = 2.5275916E-02
Cp TMSEP = 2.5275916E-02
MSE TMSEP = 2.5275916E-02

TMSEPs FOR RUN 5

1	1.1425762E-02
2	1.2566670E-02
3	1.2761764E-02
4	1.3455779E-02
5	1.4553362E-02
6	1.5473412E-02
7	1.5937347E-02
8	1.6735807E-02
9	1.7853793E-02
10	1.9062478E-02
11	2.1006519E-02
12	2.5275916E-02
13	2.7103966E-02
14	2.7148893E-02
15	2.9038427E-02

run 6

Sp TMSEP = 1.5230293E-02
Cp TMSEP = 1.5230293E-02
MSE TMSEP = 1.5230293E-02

```
TMSEPs FOR RUN            6

           1  1.1189967E-02
           2  1.1914390E-02
           3  1.2381560E-02
           4  1.2513387E-02
           5  1.3219284E-02
           6  1.3444155E-02
           7  1.5230293E-02
           8  1.5570977E-02
           9  1.5827768E-02
          10  1.6145628E-02
          11  1.6554100E-02
          12  1.6587460E-02
          13  1.6841978E-02
          14  1.7723817E-02
          15  2.0091623E-02
run                   7


 Sp  TMSEP  =      8.4279571E-03
 Cp  TMSEP  =      8.4279571E-03
 MSE TMSEP  =      8.4279571E-03


TMSEPs FOR RUN            7

           1  8.4279571E-03
           2  8.6327204E-03
           3  8.6681442E-03
           4  8.7661548E-03
           5  8.7986309E-03
           6  9.2602726E-03
           7  9.5974831E-03
           8  1.0461051E-02
           9  1.0792202E-02
          10  1.0847315E-02
          11  1.0878014E-02
          12  1.0926300E-02
          13  1.1085336E-02
          14  1.1660442E-02
          15  3.0683922E-02
run                   8


 Sp  TMSEP  =      1.9174700E-03
 Cp  TMSEP  =      1.9174700E-03
 MSE TMSEP  =      1.0391608E-02
```

TMSEPs FOR RUN 8

1	1.9174700E-03
2	2.1675059E-03
3	2.4701473E-03
4	5.9353570E-03
5	8.1103193E-03
6	1.0391608E-02
7	1.0793708E-02
8	1.2050519E-02
9	1.2089508E-02
10	1.2366496E-02
11	1.2528760E-02
12	1.2863966E-02
13	1.6300550E-02
14	1.7629199E-02
15	1.9998444E-02

run 9

Sp TMSEP = 2.9792758E-03
Cp TMSEP = 2.9792758E-03
MSE TMSEP = 2.9792758E-03

TMSEPs FOR RUN 9

1	2.5938307E-03
2	2.8756007E-03
3	2.9792758E-03
4	3.0860456E-03
5	3.8913097E-03
6	4.2111641E-03
7	4.4655493E-03
8	4.7994438E-03
9	5.4206038E-03
10	7.9978053E-03
11	8.4198406E-03
12	9.3529476E-03
13	9.8962309E-03
14	1.1325269E-02
15	1.5649561E-02

run 10

Sp TMSEP = 3.4058213E-03
Cp TMSEP = 3.4058213E-03
MSE TMSEP = 9.9750860E-03

TMSEPs FOR RUN 10

1	2.1047422E-03
2	2.5313916E-03
3	3.4058213E-03
4	4.1502793E-03
5	4.9941130E-03
6	9.2351809E-03
7	9.4529847E-03
8	9.9750860E-03
9	1.0221747E-02
10	1.0269716E-02
11	1.1170504E-02
12	1.4728523E-02
13	1.6914524E-02
14	1.9482933E-02
15	2.2141431E-02

run 11

Sp TMSEP = 1.8264394E-02
Cp TMSEP = 1.8264394E-02
MSE TMSEP = 2.6563916E-02

TMSEPs FOR RUN 11

1	5.1681511E-03
2	6.1819288E-03
3	6.7232223E-03
4	9.9291867E-03
5	1.0711449E-02
6	1.0777702E-02
7	1.1229556E-02
8	1.3848314E-02
9	1.7541861E-02
10	1.8264394E-02
11	1.8279061E-02
12	1.9208413E-02
13	2.3745688E-02
14	2.6563916E-02
15	2.8645089E-02

run 12

Sp TMSEP = 6.1405003E-03
Cp TMSEP = 6.1405003E-03
MSE TMSEP = 1.3503841E-02

TMSEPs FOR RUN 12

```
          1   1.8639894E-03
          2   1.9315871E-03
          3   2.0497970E-03
          4   2.2980263E-03
          5   2.4225425E-03
          6   2.7120986E-03
          7   6.1405003E-03
          8   7.1428162E-03
          9   7.2628092E-03
         10   8.3331009E-03
         11   8.9695957E-03
         12   1.3503841E-02
         13   1.4081514E-02
         14   1.4754524E-02
         15   1.5624065E-02
run                    13


 Sp  TMSEP  =      3.2588169E-02
 Cp  TMSEP  =      3.2588169E-02
 MSE TMSEP  =      4.0740110E-02


TMSEPs FOR RUN           13

          1   7.9008838E-04
          2   2.3885965E-03
          3   4.2190668E-03
          4   4.7011017E-03
          5   6.7453287E-03
          6   8.6223893E-03
          7   9.0424791E-03
          8   1.4171562E-02
          9   1.5555089E-02
         10   1.6289905E-02
         11   2.1787029E-02
         12   3.2588169E-02
         13   3.4396078E-02
         14   4.0740110E-02
         15   4.3201827E-02
run                    14


 Sp  TMSEP  =      1.1849147E-02
 Cp  TMSEP  =      1.1849147E-02
 MSE TMSEP  =      1.1849147E-02


TMSEPs FOR RUN           14
```

```
            1   4.8623332E-03
            2   4.9072048E-03
            3   5.1156250E-03
            4   5.8454480E-03
            5   7.1764030E-03
            6   7.8983689E-03
            7   8.7281801E-03
            8   1.1528674E-02
            9   1.1675153E-02
           10   1.1849147E-02
           11   1.2552307E-02
           12   1.3260029E-02
           13   1.5406845E-02
           14   1.6281059E-02
           15   1.9996813E-02
run                    15

 Sp  TMSEP  =      7.4793403E-03
 Cp  TMSEP  =      1.7116755E-02
 MSE TMSEP  =      1.7116755E-02

TMSEPs FOR RUN           15

            1   6.6161933E-03
            2   6.8892450E-03
            3   7.4793408E-03
            4   7.7769053E-03
            5   8.7918332E-03
            6   9.0924781E-03
            7   9.1773802E-03
            8   9.7412989E-03
            9   9.8114768E-03
           10   1.1472788E-02
           11   1.7001675E-02
           12   1.7116755E-02
           13   1.7942483E-02
           14   1.8224856E-02
           15   1.9186204E-02
run                    16

 Sp  TMSEP  =      2.4616949E-02
 Cp  TMSEP  =      2.4616949E-02
 MSE TMSEP  =      2.4616949E-02

TMSEPs FOR RUN           16
```

```
           1   2.5192862E-03
           2   6.2425360E-03
           3   6.7220321E-03
           4   8.6688502E-03
           5   1.0253071E-02
           6   1.1382802E-02
           7   1.3577277E-02
           8   1.5074901E-02
           9   1.5540805E-02
          10   1.5885668E-02
          11   1.7365742E-02
          12   2.4616949E-02
          13   2.6280256E-02
          14   2.8023722E-02
          15   2.9615028E-02
run                    17


 Sp  TMSEP  =       3.4056220E-02
 Cp  TMSEP  =       3.4056220E-02
 MSE TMSEP  =       3.4056220E-02


TMSEPs FOR RUN             17

           1   2.3214938E-03
           2   2.5332391E-03
           3   2.6155079E-03
           4   3.1588618E-03
           5   7.8641446E-03
           6   1.2681485E-02
           7   1.5965864E-02
           8   2.8908700E-02
           9   3.0276716E-02
          10   3.2774188E-02
          11   3.3868384E-02
          12   3.4056220E-02
          13   3.9103031E-02
          14   4.3489311E-02
          15   4.7426648E-02
run                    18


 Sp  TMSEP  =       1.1656721E-02
 Cp  TMSEP  =       1.1656721E-02
 MSE TMSEP  =       1.1656721E-02


TMSEPs FOR RUN             18
```

```
           1   8.8144913E-03
           2   9.2402538E-03
           3   9.3459627E-03
           4   1.0430153E-02
           5   1.0963820E-02
           6   1.1204778E-02
           7   1.1656721E-02
           8   1.3512224E-02
           9   1.3699356E-02
          10   1.5192608E-02
          11   1.5793409E-02
          12   1.6079456E-02
          13   1.7682591E-02
          14   1.8760057E-02
          15   2.5110584E-02
run                    19


 Sp  TMSEP  =       2.8525121E-03
 Cp  TMSEP  =       2.8525121E-03
 MSE TMSEP  =       8.1014968E-03


TMSEPs FOR RUN            19

           1   2.8525121E-03
           2   4.4923639E-03
           3   4.9251076E-03
           4   4.9598929E-03
           5   8.1014968E-03
           6   8.1187468E-03
           7   8.4805330E-03
           8   8.7108472E-03
           9   8.9461999E-03
          10   1.0312763E-02
          11   1.2119596E-02
          12   1.3127306E-02
          13   1.3978862E-02
          14   1.6505282E-02
          15   2.0519499E-02
run                    20


 Sp  TMSEP  =       3.8725424E-03
 Cp  TMSEP  =       3.8725424E-03
 MSE TMSEP  =       2.3732318E-02


TMSEPs FOR RUN            20
```

```
          1   2.5775689E-03
          2   2.8687967E-03
          3   3.0975479E-03
          4   3.8725424E-03
          5   5.0329203E-03
          6   7.2322008E-03
          7   8.9591537E-03
          8   9.2313904E-03
          9   1.3217278E-02
         10   1.6817946E-02
         11   1.7074332E-02
         12   1.9801777E-02
         13   2.1266723E-02
         14   2.3732318E-02
         15   2.6426652E-02
run                    21


 Sp  TMSEP  =       1.7052114E-02
 Cp  TMSEP  =       1.7052114E-02
 MSE TMSEP  =       1.7052114E-02


TMSEPs FOR RUN            21

          1   1.7052114E-02
          2   2.2108998E-02
          3   2.2773303E-02
          4   2.3853436E-02
          5   2.3881707E-02
          6   2.5029171E-02
          7   2.5230104E-02
          8   2.6611436E-02
          9   2.6725085E-02
         10   2.7688080E-02
         11   2.7753545E-02
         12   2.9282084E-02
         13   2.9369567E-02
         14   2.9715974E-02
         15   3.1632055E-02
run                    22


 Sp  TMSEP  =       5.5181938E-03
 Cp  TMSEP  =       5.5181938E-03
 MSE TMSEP  =       5.5181938E-03


TMSEPs FOR RUN            22
```

```
           1   3.6380207E-03
           2   3.9318283E-03
           3   3.9480580E-03
           4   4.1614547E-03
           5   4.2129597E-03
           6   4.3571047E-03
           7   5.5181938E-03
           8   5.7162573E-03
           9   5.8647827E-03
          10   6.0956995E-03
          11   6.8830424E-03
          12   7.0184362E-03
          13   7.4183247E-03
          14   7.5743883E-03
          15   1.9468121E-02
run                    23


 Sp  TMSEP  =       1.2474651E-02
 Cp  TMSEP  =       1.2474651E-02
 MSE TMSEP  =       1.6586190E-02


TMSEPs FOR RUN             23

           1   1.7154425E-03
           2   1.8433610E-03
           3   2.1247640E-03
           4   3.7305583E-03
           5   7.3689125E-03
           6   7.6402198E-03
           7   1.0252496E-02
           8   1.0763265E-02
           9   1.1819728E-02
          10   1.2474651E-02
          11   1.3044675E-02
          12   1.5219216E-02
          13   1.6586190E-02
          14   1.7523689E-02
          15   1.8898807E-02
run                    24


 Sp  TMSEP  =       2.0092988E-02
 Cp  TMSEP  =       2.0092988E-02
 MSE TMSEP  =       2.0092988E-02


TMSEPs FOR RUN             24
```

```
          1   3.4344932E-03
          2   6.0286978E-03
          3   6.0683978E-03
          4   6.0691400E-03
          5   6.2191375E-03
          6   7.2317598E-03
          7   8.2883285E-03
          8   1.1216477E-02
          9   1.4045537E-02
         10   1.4915518E-02
         11   1.5253766E-02
         12   1.5417388E-02
         13   1.9777540E-02
         14   2.0092988E-02
         15   2.1272380E-02
run                    25


 Sp  TMSEP  =      2.0874506E-03
 Cp  TMSEP  =      2.0874506E-03
 MSE TMSEP  =      2.0009512E-02


TMSEPs FOR RUN           25

          1   2.0874506E-03
          2   3.2037264E-03
          3   3.9723651E-03
          4   4.2093988E-03
          5   5.0320607E-03
          6   5.2119563E-03
          7   8.8053672E-03
          8   1.0451779E-02
          9   1.2837918E-02
         10   1.3056729E-02
         11   1.4630186E-02
         12   2.0009512E-02
         13   2.1082407E-02
         14   2.1697044E-02
         15   2.2916408E-02
run                    26


 Sp  TMSEP  =      9.8276651E-04
 Cp  TMSEP  =      9.8276651E-04
 MSE TMSEP  =      9.8276651E-04


TMSEPs FOR RUN           26
```

```
            1   3.6615500E-04
            2   4.0323741E-04
            3   9.8276651E-04
            4   1.0234589E-03
            5   1.0259064E-03
            6   1.0351705E-03
            7   1.1282420E-03
            8   1.2075821E-03
            9   1.2196909E-03
           10   1.3010738E-03
           11   1.5202522E-03
           12   1.5768777E-03
           13   1.7284348E-03
           14   1.9685118E-03
           15   1.2083132E-02
run                    27


 Sp  TMSEP  =       6.8846606E-03
 Cp  TMSEP  =       6.8846606E-03
 MSE TMSEP  =       1.9304380E-02


TMSEPs FOR RUN            27

            1   6.8846606E-03
            2   7.2969943E-03
            3   9.8057995E-03
            4   1.0382390E-02
            5   1.2498997E-02
            6   1.3136022E-02
            7   1.3230429E-02
            8   1.3853890E-02
            9   1.3952672E-02
           10   1.4731946E-02
           11   1.7043175E-02
           12   1.9304380E-02
           13   2.0270826E-02
           14   2.0551469E-02
           15   2.1667840E-02
run                    28


 Sp  TMSEP  =       3.7091917E-03
 Cp  TMSEP  =       3.7091917E-03
 MSE TMSEP  =       3.7091917E-03


TMSEPs FOR RUN            28
```

```
          1   3.5498959E-03
          2   3.7091917E-03
          3   3.8102050E-03
          4   3.9531672E-03
          5   4.0174657E-03
          6   4.0629935E-03
          7   4.0745367E-03
          8   4.0909247E-03
          9   4.2000455E-03
         10   4.2432775E-03
         11   4.3620034E-03
         12   4.4159661E-03
         13   4.4462909E-03
         14   4.5218416E-03
         15   1.1852265E-02
run                    29


 Sp  TMSEP  =      8.8697784E-03
 Cp  TMSEP  =      8.8697784E-03
 MSE TMSEP  =      8.8697784E-03


TMSEPs FOR RUN           29

          1   8.1481282E-03
          2   8.2919262E-03
          3   8.8697784E-03
          4   8.9680441E-03
          5   9.1282884E-03
          6   9.2851929E-03
          7   9.3137734E-03
          8   9.5947059E-03
          9   9.8677045E-03
         10   9.9192383E-03
         11   1.0009223E-02
         12   1.0221915E-02
         13   1.0479798E-02
         14   1.0976000E-02
         15   1.7647961E-02
run                    30


 Sp  TMSEP  =      3.5457979E-03
 Cp  TMSEP  =      3.5457979E-03
 MSE TMSEP  =      1.3879534E-02


TMSEPs FOR RUN           30
```

```
            1   3.5457979E-03
            2   4.6322457E-03
            3   4.8427694E-03
            4   5.4951324E-03
            5   6.3336785E-03
            6   6.3531036E-03
            7   7.4292999E-03
            8   8.7361382E-03
            9   9.9864267E-03
           10   1.1301067E-02
           11   1.3879534E-02
           12   1.4671805E-02
           13   1.4762038E-02
           14   1.5681023E-02
           15   2.0051859E-02
run                    31


 Sp  TMSEP  =       1.9329343E-02
 Cp  TMSEP  =       1.9329343E-02
 MSE TMSEP  =       2.2040404E-02


TMSEPs FOR RUN            31

            1   1.3012470E-03
            2   1.3328010E-03
            3   1.6352949E-03
            4   1.7551928E-03
            5   1.8518085E-03
            6   2.0177322E-03
            7   4.5819799E-03
            8   4.9842852E-03
            9   7.8484677E-03
           10   1.1772109E-02
           11   1.2443145E-02
           12   1.9329343E-02
           13   2.1689299E-02
           14   2.2040404E-02
           15   2.4905885E-02
run                    32


 Sp  TMSEP  =       5.0999749E-02
 Cp  TMSEP  =       5.6429233E-02
 MSE TMSEP  =       6.2712573E-02


TMSEPs FOR RUN            32
```

```
           1   7.7922344E-03
           2   9.5617203E-03
           3   9.6836761E-03
           4   1.1141410E-02
           5   1.4167450E-02
           6   1.4421574E-02
           7   1.5424208E-02
           8   1.6907291E-02
           9   2.9010303E-02
          10   2.9186631E-02
          11   3.2019943E-02
          12   5.0999749E-02
          13   5.5162951E-02
          14   5.6429233E-02
          15   6.2712573E-02
run                    33


 Sp  TMSEP  =      8.4313720E-02
 Cp  TMSEP  =      8.4313720E-02
 MSE TMSEP  =      8.4313720E-02


TMSEPs FOR RUN           33

           1   1.6472956E-02
           2   1.9875426E-02
           3   2.3888426E-02
           4   2.4913736E-02
           5   2.6807437E-02
           6   3.0863002E-02
           7   3.6406118E-02
           8   4.1674461E-02
           9   4.9072534E-02
          10   5.0164599E-02
          11   5.6148276E-02
          12   7.2232231E-02
          13   7.6316334E-02
          14   8.4313720E-02
          15   9.0484820E-02
run                    34


 Sp  TMSEP  =      1.9720267E-03
 Cp  TMSEP  =      1.9720267E-03
 MSE TMSEP  =      1.9720267E-03


TMSEPs FOR RUN           34
```

```
            1   1.4564074E-03
            2   1.5348943E-03
            3   1.5788405E-03
            4   1.7205393E-03
            5   1.9720267E-03
            6   3.0269641E-03
            7   3.1256014E-03
            8   3.4055866E-03
            9   3.6278572E-03
           10   3.9648265E-03
           11   4.2102486E-03
           12   5.5269981E-03
           13   6.0082953E-03
           14   6.6690734E-03
           15   1.7902035E-02
run                    35


 Sp  TMSEP  =      2.2126485E-02
 Cp  TMSEP  =      2.2126485E-02
 MSE TMSEP  =      2.8070798E-02


TMSEPs FOR RUN            35

            1   2.1942048E-03
            2   3.8941863E-03
            3   5.9093987E-03
            4   5.9594219E-03
            5   1.0993944E-02
            6   1.1843625E-02
            7   1.2036902E-02
            8   1.3240814E-02
            9   1.3524123E-02
           10   1.6991993E-02
           11   1.8741287E-02
           12   1.9171823E-02
           13   2.2126485E-02
           14   2.3006901E-02
           15   2.8070798E-02
run                    36


 Sp  TMSEP  =      1.5084335E-02
 Cp  TMSEP  =      1.5084335E-02
 MSE TMSEP  =      2.3686530E-02


TMSEPs FOR RUN            36
```

```
           1   1.5084335E-02
           2   1.5954880E-02
           3   1.7274419E-02
           4   1.7573619E-02
           5   1.8410455E-02
           6   1.8481053E-02
           7   1.9330112E-02
           8   2.0087145E-02
           9   2.0443222E-02
          10   2.1288132E-02
          11   2.1304168E-02
          12   2.3686530E-02
          13   2.5245925E-02
          14   2.5533320E-02
          15   2.7481722E-02
run                    37


 Sp  TMSEP  =      3.8605933E-03
 Cp  TMSEP  =      2.6786476E-02
 MSE TMSEP  =      2.6786476E-02


TMSEPs FOR RUN           37

           1   7.0788513E-04
           2   7.3154463E-04
           3   1.3002879E-03
           4   2.6231960E-03
           5   2.9667136E-03
           6   2.9975821E-03
           7   3.8605933E-03
           8   8.0252122E-03
           9   1.6348951E-02
          10   1.8301096E-02
          11   2.0436686E-02
          12   2.2819119E-02
          13   2.6190016E-02
          14   2.6786476E-02
          15   3.0853484E-02
run                    38


 Sp  TMSEP  =      1.0595867E-02
 Cp  TMSEP  =      1.0595867E-02
 MSE TMSEP  =      2.3941863E-02


TMSEPs FOR RUN           38
```

```
          1   2.9883636E-03
          2   4.1764169E-03
          3   4.8778988E-03
          4   9.7571025E-03
          5   9.9973157E-03
          6   1.0465559E-02
          7   1.0595867E-02
          8   1.4187485E-02
          9   1.7074063E-02
         10   2.0584134E-02
         11   2.3294877E-02
         12   2.3941863E-02
         13   2.5516720E-02
         14   2.5926596E-02
         15   2.7840657E-02
run                   39


 Sp  TMSEP  =      1.0622767E-02
 Cp  TMSEP  =      1.0622767E-02
 MSE TMSEP  =      1.0622767E-02


TMSEPs FOR RUN           39

          1   1.1929536E-03
          2   1.3373954E-03
          3   1.6708911E-03
          4   3.4250987E-03
          5   3.5063413E-03
          6   4.6463534E-03
          7   4.6803518E-03
          8   8.6782407E-03
          9   8.7229246E-03
         10   8.8040661E-03
         11   1.0622767E-02
         12   1.2146195E-02
         13   1.2587945E-02
         14   1.3528628E-02
         15   1.7254293E-02
run                   40


 Sp  TMSEP  =      3.5547487E-02
 Cp  TMSEP  =      3.5547487E-02
 MSE TMSEP  =      4.1834652E-02


TMSEPs FOR RUN           40
```

```
            1   7.5619790E-04
            2   1.5781120E-03
            3   2.6893581E-03
            4   2.8271538E-03
            5   1.0157598E-02
            6   1.0704858E-02
            7   1.1176147E-02
            8   1.2656823E-02
            9   2.0222088E-02
           10   2.0667935E-02
           11   2.1879649E-02
           12   2.8531680E-02
           13   3.0591035E-02
           14   3.5547487E-02
           15   4.1834652E-02
run                    41


 Sp  TMSEP  =       2.2413865E-02
 Cp  TMSEP  =       2.2413865E-02
 MSE TMSEP  =       2.8072061E-02


TMSEPs FOR RUN            41

            1   1.0048087E-02
            2   1.5616417E-02
            3   1.5880233E-02
            4   1.6917937E-02
            5   1.7490519E-02
            6   1.7648792E-02
            7   1.8472508E-02
            8   1.9556798E-02
            9   2.0597108E-02
           10   2.1678284E-02
           11   2.2413865E-02
           12   2.4091654E-02
           13   2.7307892E-02
           14   2.8072061E-02
           15   3.0794047E-02
run                    42


 Sp  TMSEP  =       4.7981236E-03
 Cp  TMSEP  =       4.7981236E-03
 MSE TMSEP  =       4.7981236E-03


TMSEPs FOR RUN            42
```

```
          1   3.4364881E-03
          2   3.5259190E-03
          3   3.6066370E-03
          4   3.9199176E-03
          5   4.0680161E-03
          6   4.1261218E-03
          7   4.7981236E-03
          8   5.2037672E-03
          9   5.3428058E-03
         10   5.6819613E-03
         11   5.7216412E-03
         12   6.0622832E-03
         13   6.0960101E-03
         14   6.4953892E-03
         15   1.5678339E-02
run                    43


 Sp  TMSEP  =       3.9310423E-03
 Cp  TMSEP  =       3.9310423E-03
 MSE TMSEP  =       3.9310423E-03


TMSEPs FOR RUN            43

          1   1.8738484E-03
          2   2.3403340E-03
          3   2.9570037E-03
          4   3.1850252E-03
          5   3.9310423E-03
          6   4.7286372E-03
          7   4.8334445E-03
          8   4.8407987E-03
          9   5.7194503E-03
         10   6.7090066E-03
         11   7.6004523E-03
         12   9.2586232E-03
         13   9.2892852E-03
         14   1.2049303E-02
         15   1.2300548E-02
run                    44


 Sp  TMSEP  =       1.2019108E-03
 Cp  TMSEP  =       1.8140119E-02
 MSE TMSEP  =       1.8140119E-02


TMSEPs FOR RUN            44
```

```
           1   3.9169341E-04
           2   5.2826642E-04
           3   1.2019108E-03
           4   1.2748841E-03
           5   3.4131568E-03
           6   6.1522294E-03
           7   6.9718626E-03
           8   7.0441552E-03
           9   7.1675945E-03
          10   7.9703350E-03
          11   8.1011681E-03
          12   8.6136796E-03
          13   1.8140119E-02
          14   1.9227074E-02
          15   2.1237271E-02
run                    45


 Sp  TMSEP  =      3.0700916E-03
 Cp  TMSEP  =      3.0700916E-03
 MSE TMSEP  =      1.1097923E-02


TMSEPs FOR RUN           45

           1   3.0700916E-03
           2   3.3025786E-03
           3   4.1044764E-03
           4   5.5306656E-03
           5   5.6966278E-03
           6   5.9962580E-03
           7   6.9032218E-03
           8   7.6960446E-03
           9   1.0958084E-02
          10   1.1097923E-02
          11   1.1505049E-02
          12   1.1665934E-02
          13   1.1997189E-02
          14   1.2202757E-02
          15   1.3101306E-02
run                    46


 Sp  TMSEP  =      2.2392911E-03
 Cp  TMSEP  =      8.1968531E-03
 MSE TMSEP  =      1.1323700E-02


TMSEPs FOR RUN           46
```

```
           1   2.2392911E-03
           2   2.4520580E-03
           3   2.9001222E-03
           4   5.2466160E-03
           5   5.2603823E-03
           6   5.3152107E-03
           7   6.3047521E-03
           8   6.6063711E-03
           9   7.9144575E-03
          10   8.1968531E-03
          11   8.2538687E-03
          12   9.0868426E-03
          13   1.1323700E-02
          14   1.3579807E-02
          15   1.5681304E-02
run                    47


 Sp  TMSEP  =      1.0096396E-02
 Cp  TMSEP  =      1.0096396E-02
 MSE TMSEP  =      1.0096396E-02


TMSEPs FOR RUN            47

           1   9.9929059E-03
           2   1.0018086E-02
           3   1.0096396E-02
           4   1.0306062E-02
           5   1.1321264E-02
           6   1.1852885E-02
           7   1.2029038E-02
           8   1.2128158E-02
           9   1.2128995E-02
          10   1.2503473E-02
          11   1.2757318E-02
          12   1.4217354E-02
          13   1.4767253E-02
          14   1.6271524E-02
          15   2.8789150E-02
run                    48


 Sp  TMSEP  =      7.3499735E-03
 Cp  TMSEP  =      7.3499735E-03
 MSE TMSEP  =      1.9433314E-02


TMSEPs FOR RUN            48
```

```
          1   7.3499735E-03
          2   7.8419475E-03
          3   1.0533275E-02
          4   1.1174344E-02
          5   1.1179664E-02
          6   1.1932554E-02
          7   1.5124455E-02
          8   1.5476356E-02
          9   1.5993901E-02
         10   1.6502503E-02
         11   1.7872537E-02
         12   1.9433314E-02
         13   2.0548729E-02
         14   2.0941224E-02
         15   2.2205124E-02
run                   49


 Sp  TMSEP  =      4.2170435E-02
 Cp  TMSEP  =      4.2170435E-02
 MSE TMSEP  =      4.2170435E-02


TMSEPs FOR RUN           49

          1   1.5701497E-02
          2   1.6772849E-02
          3   1.7934496E-02
          4   1.9688854E-02
          5   2.1170381E-02
          6   2.2765471E-02
          7   2.4873454E-02
          8   2.5819845E-02
          9   2.7326189E-02
         10   2.9126355E-02
         11   3.0813400E-02
         12   3.1698938E-02
         13   3.3878606E-02
         14   4.2170435E-02
         15   4.4911824E-02
run                   50


 Sp  TMSEP  =      2.2768520E-02
 Cp  TMSEP  =      2.2768520E-02
 MSE TMSEP  =      3.6368068E-02


TMSEPs FOR RUN           50
```

```
           1   1.5540421E-02
           2   1.6995009E-02
           3   1.7393902E-02
           4   2.0387221E-02
           5   2.1287283E-02
           6   2.1346433E-02
           7   2.2153363E-02
           8   2.2768520E-02
           9   2.3359068E-02
          10   2.4774464E-02
          11   2.5770180E-02
          12   2.8919019E-02
          13   3.0525209E-02
          14   3.6368068E-02
          15   3.8798217E-02
run                    51


 Sp  TMSEP  =       2.4604530E-03
 Cp  TMSEP  =       2.4604530E-03
 MSE TMSEP  =       7.0150704E-03


TMSEPs FOR RUN            51

           1   1.6393900E-03
           2   1.8884840E-03
           3   2.2606149E-03
           4   2.4604530E-03
           5   2.4627568E-03
           6   2.6644068E-03
           7   2.7719922E-03
           8   3.1818897E-03
           9   3.1976223E-03
          10   3.4056555E-03
          11   7.0150704E-03
          12   7.1953204E-03
          13   7.7710031E-03
          14   7.9887751E-03
          15   2.0532234E-02
run                    52


 Sp  TMSEP  =       1.5042799E-02
 Cp  TMSEP  =       3.2231480E-02
 MSE TMSEP  =       3.2231480E-02


TMSEPs FOR RUN            52
```

```
            1   1.3510313E-02
            2   1.4268238E-02
            3   1.5042799E-02
            4   1.5315600E-02
            5   1.6465578E-02
            6   1.7620863E-02
            7   1.8399971E-02
            8   1.8756019E-02
            9   2.1330154E-02
           10   2.4062017E-02
           11   2.4590410E-02
           12   2.6639560E-02
           13   2.7676962E-02
           14   3.2231480E-02
           15   3.5307933E-02
run                    53


 Sp  TMSEP  =       4.1374839E-03
 Cp  TMSEP  =       4.1374839E-03
 MSE TMSEP  =       4.1374839E-03


TMSEPs FOR RUN             53

            1   1.4167477E-03
            2   1.8064859E-03
            3   1.8127215E-03
            4   4.1374839E-03
            5   5.3663515E-03
            6   5.3790277E-03
            7   5.4524424E-03
            8   5.7910713E-03
            9   6.2028468E-03
           10   6.9525018E-03
           11   8.6443033E-03
           12   8.7333312E-03
           13   9.9164629E-03
           14   1.0268040E-02
           15   1.1267337E-02
run                    54


 Sp  TMSEP  =       1.4347644E-02
 Cp  TMSEP  =       1.4347644E-02
 MSE TMSEP  =       1.4347644E-02


TMSEPs FOR RUN             54
```

```
            1   5.8415225E-03
            2   6.4000203E-03
            3   7.0034480E-03
            4   8.8544022E-03
            5   9.1899643E-03
            6   9.4587682E-03
            7   9.9534485E-03
            8   1.0385258E-02
            9   1.0859730E-02
           10   1.1279927E-02
           11   1.4146212E-02
           12   1.4347644E-02
           13   1.5366105E-02
           14   1.7616469E-02
           15   1.8684378E-02
run                    55


 Sp  TMSEP  =       1.2993388E-02
 Cp  TMSEP  =       1.2993388E-02
 MSE TMSEP  =       1.9165561E-02


TMSEPs FOR RUN            55

            1   4.9696909E-03
            2   7.7763824E-03
            3   9.6101258E-03
            4   1.1696715E-02
            5   1.2170565E-02
            6   1.2716134E-02
            7   1.2899847E-02
            8   1.2993388E-02
            9   1.5015327E-02
           10   1.5555417E-02
           11   1.5641246E-02
           12   1.6055508E-02
           13   1.7838184E-02
           14   1.9165561E-02
           15   2.1346293E-02
run                    56


 Sp  TMSEP  =       6.3008070E-02
 Cp  TMSEP  =       6.3008070E-02
 MSE TMSEP  =       7.3209882E-02


TMSEPs FOR RUN            56
```

```
           1   6.1677363E-02
           2   6.2864579E-02
           3   6.3008070E-02
           4   6.6310331E-02
           5   6.6765495E-02
           6   6.6853113E-02
           7   7.1422204E-02
           8   7.1454287E-02
           9   7.3190719E-02
          10   7.3209882E-02
          11   7.5800024E-02
          12   7.7444568E-02
          13   7.8586839E-02
          14   7.9587929E-02
          15   8.5475355E-02
run                    57


 Sp  TMSEP  =       3.7371186E-03
 Cp  TMSEP  =       3.7371186E-03
 MSE TMSEP  =       3.7371186E-03


TMSEPs FOR RUN             57

           1   3.7371186E-03
           2   3.8739347E-03
           3   3.9046754E-03
           4   4.0383120E-03
           5   4.1861599E-03
           6   4.4727097E-03
           7   4.6059825E-03
           8   4.7026719E-03
           9   5.1347492E-03
          10   5.8260681E-03
          11   5.8471588E-03
          12   6.2553254E-03
          13   6.4490214E-03
          14   7.0934095E-03
          15   9.5695062E-03
run                    58


 Sp  TMSEP  =       1.2361378E-02
 Cp  TMSEP  =       1.2361378E-02
 MSE TMSEP  =       1.2361378E-02


TMSEPs FOR RUN             58
```

```
          1   2.4738992E-03
          2   3.7456378E-03
          3   3.8988995E-03
          4   4.2070122E-03
          5   4.5560552E-03
          6   4.8246444E-03
          7   6.3058720E-03
          8   1.2361378E-02
          9   1.3068646E-02
         10   1.3421705E-02
         11   1.3535345E-02
         12   1.4164726E-02
         13   1.6804852E-02
         14   1.7755972E-02
         15   2.3304023E-02
run                    59

 Sp  TMSEP  =      1.5110006E-02
 Cp  TMSEP  =      1.5110006E-02
 MSE TMSEP  =      1.5110006E-02


TMSEPs FOR RUN           59

          1   1.0106159E-02
          2   1.1130731E-02
          3   1.1358641E-02
          4   1.1672366E-02
          5   1.2471750E-02
          6   1.3456321E-02
          7   1.3687690E-02
          8   1.5110006E-02
          9   1.5243942E-02
         10   1.7136147E-02
         11   1.8399373E-02
         12   1.8740477E-02
         13   1.9748218E-02
         14   2.0658856E-02
         15   2.3248075E-02


run                    60

 Sp  TMSEP  =      1.8573448E-02
 Cp  TMSEP  =      1.8573448E-02
 MSE TMSEP  =      1.8573448E-02
```

TMSEPs FOR RUN 60

1	1.1594909E-02
2	1.2438751E-02
3	1.2594829E-02
4	1.2842594E-02
5	1.3457467E-02
6	1.3940695E-02
7	1.5752792E-02
8	1.8573448E-02
9	2.0086005E-02
10	2.2414638E-02
11	2.3428341E-02
12	2.4330338E-02
13	2.5231192E-02
14	2.5254766E-02
15	2.7355636E-02

Bibliography

1. Barr, David R. "Interpretations of Mallows C_p Criterion in all Possible Regression Models." Unpublished report. AFIT/ENC, Wright-Patterson AFB OH, 1-7.

2. Berk, Kenneth N. "Comparing Subset Regression Procedures," *Technometrics, 20*: 1-6 (February 1978).

3. Breiman, L. and D. Freedman. "How Many Variables Should Be Entered in a Regression Equation?" *Journal of the American Statistical Association, 78*: 131-6 (March 1983).

4. Cafarella, 2Lt Joseph R., Jr. *Cross Validation of Selection of Variables in Multiple Regression*, MS Thesis, AFIT/GOR/MA/79D-2. School of Engineering, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, December 1979.

5. Dempster, A.P., Martin Schatzoff, and Nammy Wermuth. "A Simulation Study of Alternatives to Ordinary Least Squares," *Journal of the American Statistical Association, 72*: 77-106 (March 1977).

6. Diehr, George, and Hoflin, R. Donald. "Approximating the Distribution of the Sample R^2 in Best Subset Regressions," *Technometrics, 16*: 317-320 (May 1974).

7. Draper, Norman, and Harry Smith. *Applied Regression Analysis, Second Edition* New York: John Wiley & Sons, 1981.

8. Flack, Virgina F., and Chang, Potter C. "Frequency of Selecting Noise Variables in Subset Regression Analysis: A Simulation Study", *The American Statistician, 41*: 84-86 (February 1987).

9. Freedman, David A. "A Note on Screening Regression Equations" *The American Statistician, 37*: 152-155 (May 1983).

10. Furnival, George M., and Robert W. Wilson Jr. "Regressions by Leaps and Bounds," *Technometrics, 16*: 499-511 (November 1974).

11. Healy, M.J. "The Use of R^2 as a Measure of Goodness of Fit" *Journal of the Royal Statistical Society Series A, 147*: 608-609 (1984).

12. Hocking, R.R. "The Analysis and Selection of Variables in Linear Regression," *Biometrika, 32*: 1-49 (March 1976).

13. Hoerl, Roger W., Arthur E. Hoerl, and John H. Schuenemeyer "A Simulation of Biased Estimation and Subset Selection Regression Techniques," *Technometrics, 24*: 370-380 (November 1986).

14. Huang, D.Y., and S. Panchapakensan. "On Eliminating Inferior Regression Models," *Commun. Statist. Theor. Method, 11*: 751-759 (1982).

15. Judge, George, W.E. Griffiths, R. Carter Hill, Helmut Lutkepohl, and Tsoung-Chao Lee. *The Theory and Practice of Econometrics, Second Edition*. New York: John Wiley & Sons, Inc., 1985.

16. Kempthorne, Peter J. "Admissible Variable-Selection Procedures When Fitting Regression Models by Least Squares for Prediction," *Biometrika, 71*: 593-7 (1983).

17. Klein, R.W., and S.J. Brown. "Model Selection When There is `Minimal' Prior Information," *Econometrica, 52*: 1291-1311 (September 1984).

18. Lowell, Michael C. "Data Mining," *The Review of Econometrics and Statistics, 65*: 1-12 (February 1983).

19. Makin, Capt James R. *An Evaluation of Ridge Regression In Cost Estimation*, MS Thesis, AFIT/GOR/OS/81D-6. School of Engineering, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, December 1981.

20. Miller, Alan J. "Selection of Subsets of Regression Variables," *Journal of the Royal Statistical Society A, 147*: 389-425 (1984).

21. Narula, Subhash C. and John F. Wellington. "Selection of Variables in Linear Regression: A Pragmatic Approach," *Journal of Statistical Computations and Simulation, 12*: 59-172 (1983).

22. Norberg, L. "On Variable Selection in Generalized Linear and Related Regression Models," *Commun. Statist.- Theor. Method, 11*: 2427-2449 (1982).

23. Oliver, V.I. "On the Relationship Between the Sample Size and the Number of Variables in a Linear Regression Model," *Journal of Statistical Computations and Simulation, A7(6)*: 509-16 (1978).

24. Pulcher, Capt Larry J. *Criterion for Selection of Variables in a Regression Analysis*. MS Thesis, AFIT/GOR/MA/78D-5. School for Engineering, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, December 1978.

25. Shibata, Ritei. "An Optimal Selection of Regression Variables," *Biometrika, 68*: 45-54 (1981).

26. Sparks, R.S., W. Zucchini, and D. Coutsourides. "On Variable Selection in Multivariate Regression," *Communication and Statistical Theory Methods, 14*: 1569-1587 (1985).

27. Thompson, Mary L. "Selection of Variables in Multiple Regression: Part I. A Review and Evaluation," *International Statistical Review, 46*: 1-19 (1978).

28. Trader, Ramona L. "A Bayesian Predictive Approach to the Selection of Variables in Multiple Regression," *Communication and Statistical Theory Methods, 12(13)*: 1553-1567 (1983).

29. Wang, P.C. "Adding a Variable in Generalized Linear Models," *Technometrics, 27*: 273-276 (August 1985).

30. Weisberg, Sanford. "A Statistic for Allocating C_p to Individual Cases," *Technometrics, 23*: 27-31 (February 1981).

VITA

First Lieutenant Ross J. Hansen was born [redacted] He graduated from high school [redacted] in 1982 and attended the United States Air Force Academy, from which he received the degree of Bachelor of Science in Operations Research in May 1986. Upon graduation, he received a commission in the USAF.

[redacted]

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release distribution unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) AFIT/GOR/MA/88D-3	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION School of Engineering 6b. OFFICE SYMBOL (*If applicable*) AFIT/ENC	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (*City, State, and ZIP Code*) Air Force Institute of Technology Wright-Patterson AFB, OH 45433-6583	7b. ADDRESS (*City, State, and ZIP Code*)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION 8b. OFFICE SYMBOL (*If applicable*)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (*City, State, and ZIP Code*)	10 SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO / WORK UNIT ACCESSION NO.
11. TITLE (*Include Security Classification*) See Box 19	
12. PERSONAL AUTHOR(S) Ross J. Hansen, B.S., 1Lt, USAF	
13a. TYPE OF REPORT MS Thesis 13b. TIME COVERED FROM _____ TO _____ 14. DATE OF REPORT (*Year, Month, Day*) 1988 December 15. PAGE COUNT 124	
16. SUPPLEMENTARY NOTATION	

17. COSATI CODES			18. SUBJECT TERMS (*Continue on reverse if necessary and identify by block number*)
FIELD	GROUP	SUB-GROUP	
12	03		Statistics Least Squares Method Regression Analysis

19. ABSTRACT (*Continue on reverse if necessary and identify by block number*)

Title: A COMPARISON OF VARIABLE SELECTION CRITERIA FOR MULTIPLE LINEAR REGRESSION: A SIMULATION STUDY

Thesis Chairman: David R. Barr, PhD
Assistant Professor of Mathematics

10 Jan 89

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL David R. Barr, PhD	22b. TELEPHONE (*Include Area Code*) (513) 255-3098 22c. OFFICE SYMBOL AFIT/ENC

UNCLASSIFIED

19. (CONTINUED) ABSTRACT

The purpose of this thesis was to identify three promising least squares selection procedures discussed in the literature during the previous decade and then test ~~them~~ using simulation. The three criteria chosen for this study were minimum mean square error (Min MSE), minimum Sp, and minimum Cp.

Most of the previous simulations in this area are limited to investigating the usefulness of variable selection criteria when all relevant regressors and some noise variables are available. It is questionable whether all relevant variables will be included. This research has examined the effects of not including a significant variable in the variable pool.

In examining each criterion, emphasis was placed on the technique's performance under varying amounts of multicollinearity, variable variation, number of variables, and sample size. Response Surface Methodology was used to determine the effects of varying these factors. A comparison was then made using the results from the Response Surface.

To supplement the simulation research a comprehensive literature review of the most current journal articles dealing with several least squares criteria has been provided. This review includes a discussion of each technique's strengths and weaknesses. Since many of the least squares variable selection criteria are addressed, this thesis serves as a useful starting place for various regression questions.